



(11) **EP 3 149 168 B1**

(12) **EUROPEAN PATENT SPECIFICATION**

(45) Date of publication and mention
of the grant of the patent:
22.09.2021 Bulletin 2021/38

(21) Application number: **15798953.4**

(22) Date of filing: **27.05.2015**

(51) Int Cl.:
C12N 15/10 (2006.01)

(86) International application number:
PCT/US2015/032760

(87) International publication number:
WO 2015/184016 (03.12.2015 Gazette 2015/48)

(54) **HIGH-THROUGHPUT ASSEMBLY OF GENETIC ELEMENTS**

HOCHDURCHSATZANORDNUNG GENETISCHER ELEMENTE

ASSEMBLAGE À HAUT RENDEMENT D'ÉLÉMENTS GÉNÉTIQUES

(84) Designated Contracting States:
**AL AT BE BG CH CY CZ DE DK EE ES FI FR GB
GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO
PL PT RO RS SE SI SK SM TR**

(30) Priority: **27.05.2014 US 201462003331 P**

(43) Date of publication of application:
05.04.2017 Bulletin 2017/14

(73) Proprietors:
• **The Broad Institute, Inc.**
Cambridge, MA 02142 (US)
• **Massachusetts Institute Of Technology**
Cambridge, MA 02139 (US)

(72) Inventors:
• **NICOL, Robert**
Cambridge, MA 02142 (US)
• **Andrews, Lauren**
Cambridge, MA 02139 (US)
• **MIKKELSEN, Tarjei**
Cambridge, MA 02139 (US)
• **VOIGT, Christopher**
Belmont, MA 02478 (US)

(74) Representative: **De Clercq, Ann G. Y. et al**
De Clercq & Partners
Edgard Gevaertdreef 10a
9830 Sint-Martens-Latem (BE)

(56) References cited:
WO-A1-2008/000090 WO-A2-2008/054543
US-A1- 2007 269 870 US-A1- 2010 124 768
US-A1- 2013 298 265 US-A1- 2014 121 116

- **SEAN C. SLEIGHT ET AL: "Randomized BioBrick Assembly: A Novel DNA Assembly Method for Randomizing and Optimizing Genetic Circuits and Metabolic Pathways", ACS SYNTHETIC BIOLOGY, vol. 2, no. 9, 20 September 2013 (2013-09-20), pages 506-518, XP055400088, Washington, DC,USA ISSN: 2161-5063, DOI: 10.1021/sb4000542**
- **HIATT JOSEPH B ET AL: "Parallel, tag-directed assembly of locally derived short sequence reads (+ Online Methods)", NATURE METHODS, vol. 7, no. 2, February 2010 (2010-02), pages 119-122+3PP, XP055005905, DOI: 10.1038/nmeth.1416**
- **LAUREN B. A. WOODRUFF ET AL: "Registry in a tube: multiplexed pools of retrievable parts for genetic design space exploration", NUCLEIC ACIDS RESEARCH, 21 December 2016 (2016-12-21), page gkw1226, XP055422597, ISSN: 0305-1048, DOI: 10.1093/nar/gkw1226**
- **MICHAEL J SMANSKI ET AL: "Functional optimization of gene clusters by combinatorial design and assembly (+ Online Methods)", NATURE BIOTECHNOLOGY, 24 November 2014 (2014-11-24), pages 1-9+3PP, XP055157548, ISSN: 1087-0156, DOI: 10.1038/nbt.3063**
- **None**

Note: Within nine months of the publication of the mention of the grant of the European patent in the European Patent Bulletin, any person may give notice to the European Patent Office of opposition to that patent, in accordance with the Implementing Regulations. Notice of opposition shall not be deemed to have been filed until the opposition fee has been paid. (Art. 99(1) European Patent Convention).

EP 3 149 168 B1

Description**CROSS REFERENCE TO RELATED APPLICATION**

- 5 **[0001]** This application claims the priority benefit of the earlier filing date of U.S. Provisional Application No. 62/003,331, filed May 27, 2014

STATEMENT OF GOVERNMENT SUPPORT

- 10 **[0002]** This invention was made with government support under Living Foundries grant HR0011-13-1-0001 awarded by US Defense Advanced Research Projects Agency (DARPA). The government has certain rights in the invention.

FIELD

- 15 **[0003]** The present disclosure relates to methods for high-throughput assembly and sequence verification of nucleic acid constructs.

BACKGROUND

- 20 **[0004]** The field of synthetic biology applies the principles of engineering and rational design to the design and construction of biological systems. DNA-encoded functional units, such as genes or regulatory elements, can be abstracted into interchangeable "parts," which can be assembled in new combinations to produce new genetic designs. For example, BioBricks™ are modular segments of DNA flanked by standardized sequences, enabling assembly of BioBrick™ concatemers by standard restriction digest/ligation methods.
- 25 **[0005]** The cost and turnaround time of DNA assembly remains comparatively high, limiting the rate at which genetic designs can be constructed and optimized. Furthermore, the assembly of large constructs of DNA (for example, 1000+ base pairs) remains an error prone process. Thus, there is a need in the art for high-throughput methods that combine efficient and inexpensive DNA assembly with sequence verification, thus permitting rapid, inexpensive prototyping and iteration of genetic designs.

SUMMARY OF THE DISCLOSURE

- 30 **[0006]** Disclosed are methods for generating a set of nucleic acid elements for the combinatorial construction of a genetic design. Embodiments of the disclosed method include: providing a plurality of nucleic acid parts, comprising one or more categories of nucleic acid parts; assembling the nucleic acid parts to form one or more nucleic acid elements, wherein the nucleic acid elements comprise at least two sequences selected from the plurality of nucleic acid parts; and assembling onto the nucleic acid elements or parts a 3' and/or a 5' flanking nucleic acid sequence that orders assembly of higher-order nucleic acid constructs, facilitates retrieval and/or identification of the nucleic acid element. In some embodiments, the nucleic acid parts further comprise one or more nucleic acid assembly sequences that order assembly.
- 35 In some embodiments, the nucleic acid assembly sequence comprises one or more restriction sites and, prior to assembly, they are cleaved to produce sticky ends to facilitate assembly. In some embodiments, the 3' and/or 5' flanking sequences comprise one or more restriction sites and, prior to assembly, they are cleaved with restriction enzymes to produce sticky or blunt ends to facilitate assembly. In some embodiments, the one or more restriction sites comprise at least one Type IIs restriction site. In some embodiments, cleavage of the Type IIs restriction site results in cohesive end overhangs at the 5' and 3' ends of the nucleic acid sequences such that the sequences are assembled into a continuous DNA fragment with the designed order and orientation of its components. In some embodiments, the nucleic acid parts further comprise one or more nucleic acid barcodes. In some embodiments, the 3' and/or 5' flanking sequences comprise one or more nucleic acid barcodes.
- 40 **[0007]** In certain embodiments, the method includes: providing a plurality of nucleic acid parts, including one or more of: a set of nucleic acid sequences comprising different promoters; a set of nucleic acid sequences comprising a set of different ribosomal binding sites; a set of nucleic acid sequences encoding a polypeptide; and/or a set of nucleic acid sequences comprising a set of different transcription terminators; assembling nucleic acid parts to form one or more nucleic acid elements, wherein the nucleic acid elements include at least two sequences selected from the plurality of parts; and determining the sequence of the nucleic acid elements (for example via nucleic acid sequencing, nucleic acid amplification, nucleic acid hybridization, or a combination thereof). In certain embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same set. In certain embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are from the same set. In some embodiments, the parts, such as the parts in the pools have variations in sequence, for example variable polypeptide or CDS sequences within a pool,
- 45
- 50
- 55

for example different orthologs and homologs of a known or predicted enzyme or other function. Alternatively, different synthetic or refactored coding sequences could be varied within the assembled pools (same polypeptide but different DNA sequences due to codon redundancy). In certain embodiments, the nucleic acid element includes a functional cistron or a subunit thereof. In certain embodiments, the nucleic acid element includes a nucleic acid sequence encoding a protein. In certain embodiments, the method further includes assembling a nucleic acid barcode to the 5' and/or 3' end of the nucleic acid element. In certain embodiments, the barcodes include a sequencing adaptor. In certain embodiments, the method further includes pooling the plurality of plasmids. In certain embodiments, the barcodes include a sequencing adaptor. In certain embodiments, the barcode further includes universal priming sites. In certain embodiments, the barcodes further include sequence specific priming sites to facilitate retrieval. In certain embodiments, each of the barcodes further includes a unique molecular identifier. In certain embodiments, each of the barcodes includes one or more indexes, one or more sequences that enable capture and/or amplification, and/or one or more sequences that enable library construction. In certain embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators are sequence variants. In certain embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators includes at least one Type IIs restriction site and, prior to assembly, they are cleaved at the Type IIs restriction site to produce sticky ends to facilitate further assembly. In certain embodiments, the method further includes amplifying the nucleic acid elements to form a mixture of amplified nucleic acid elements, for example via polymerase chain reaction (PCR). In some examples, the method further includes inserting the nucleic acid elements into nucleic acid vectors, thereby generating a plurality of plasmids containing the nucleic acid elements. In certain embodiments, the method further includes pooling the plurality of plasmids. In certain embodiments, the method further includes determining the sequence of the nucleic acid elements via nucleic acid sequencing, nucleic acid amplification, nucleic acid hybridization, or a combination thereof. In certain embodiments, the method further includes retrieving one or more of the sequence-verified nucleic acid elements, for example from a pool of such elements.

[0008] Also disclosed is a method of generating a pool of higher-order nucleic acid constructs. The method including: retrieving two or more nucleic acid elements from the sequence-verified pool of nucleic acid elements; assembling the two or more nucleic acid elements to form one or more higher-order nucleic acid constructs; and determining the sequence of the higher-order nucleic acid constructs, thereby generating a sequence-verified pool of higher-order nucleic acid constructs. In certain embodiments, the method further includes assembling a nucleic acid barcode to the 5' and/or 3' end of the higher-order nucleic acid construct. In certain embodiments, the method further includes determining the sequence of the higher-order nucleic acid constructs via nucleic acid sequencing, nucleic acid amplification, nucleic acid hybridization, or a combination thereof. In certain embodiments, the method further includes amplifying the higher-order nucleic acid constructs to form a mixture of amplified, higher-order nucleic acid constructs. In certain embodiments, the higher-order nucleic acid construct includes two or more nucleic acid sequences encoding a protein. In certain embodiments, the method further includes inserting the higher-order nucleic acid constructs into a nucleic acid vectors, thereby generating a plurality of plasmids containing the higher-order nucleic acid constructs.

[0009] Disclosed is a set of nucleic acid elements for the combinatorial construction of genetic elements or amplification products thereof. In some embodiments the elements include one or more nucleic acid elements assembled from: a set of nucleic acid sequences comprising different promoters; a set of nucleic acid sequences comprising a set of different ribosomal binding sites; a set of nucleic acid sequences encoding a polypeptide; and/or a set of nucleic acid sequences comprising a set of different transcription terminators, wherein sequence of the nucleic acid element is determined after assembly. In some embodiments, the sequence is verified prior to assembly. In some embodiments, the sequence is verified after assembly. In certain embodiments, the nucleic acid elements include at least three sequences selected from the set of a promoters, ribosomal binding sites, coding sequences, and transcription terminators. In certain embodiments, the nucleic acid elements include at least four sequences selected from the set of promoters, ribosomal binding sites, coding sequences and transcription terminators. In certain embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same set. In certain embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are from the same set. In certain embodiments, the nucleic acid element includes a functional cistron or a subunit thereof. In certain embodiments, the nucleic acid element includes a nucleic acid encoding a protein. In certain embodiments, the sequence-verified nucleic acid elements further include a nucleic acid barcode to the 5' and/or 3' end of the nucleic acid element. In certain embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the

set of different transcription terminators are sequence variants. Also disclosed are a set of plasmids including the set of sequence-verified nucleic acid elements.

[0010] Further disclosed is a method of optimizing a combination of nucleic acid elements for the production of an optimum desired phenotype, the method including: retrieving a plurality of sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs; assembling the retrieved sequences to construct a biological pathway assembly; testing each of the assembled biological pathway assemblies for the desired phenotype; determining which characteristic or characteristics of the assembled biological pathway contributes to the desired phenotype by comparing multiple assemblies; and based on the determination, retrieving a plurality of different sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs. In one embodiment, the method further includes assembling the retrieved sequences to construct a biological pathway assembly; testing each of the of the assembled biological pathway assemblies for the desired phenotype; and repeating until a biological pathway assembly is identified that produces said optimum desired phenotype.

[0011] Further disclosed is a pool of higher-order nucleic acid constructs or amplification products thereof, including one or more nucleic acid elements. In certain embodiments, the higher-order nucleic acid constructs further include a nucleic acid barcode on the 5' and/or 3' end.

[0012] Further disclosed is a kit including a pool of sequence-verified nucleic acid elements and/or a pool of higher-order nucleic acid constructs; and a plurality of primers for retrieving one or more sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs.

[0013] Further disclosed are one or more electronic files including sequence information for a pool of sequence-verified nucleic acid elements and/or a pool of higher-order nucleic acid constructs.

[0014] Further disclosed is a software package that can implement a desired design as a set of oligos for automated retrieval. The software may be guided by design of experiments algorithms (DOE), metabolic flux or protein engineering models, or genetic circuit design packages. The software may also be guided by combinatorial design algorithms.

BRIEF DESCRIPTION OF THE DRAWINGS

[0015]

Figure 1 is a diagram showing exemplary methods that enable the engineering of complex genetic systems on a realizable timeframe, such as through rapid, targeted combinatorial searches of relevant design space.

Figure 2 is a diagram showing components of nitrogenase biosynthesis and nitrogen fixation pathways.

Figure 3 is a set of diagrams showing the results of 96 whole cluster sequence analysis to identify error modes in existing assembly methodologies.

Figure 4 is a diagram showing the assembly of a verified nif cluster using 53 iterations of assembly reactions, transformations, colony selection, purification, and normalization; and 64 Sanger sequencing reactions.

Figure 5 is a flowchart illustrating a method of multiplex scarless DNA assembly and sequencing followed by tag-directed retrieval.

Figure 6 is a diagram showing the steps of a method for barcoded multiplex assembly and tag-directed retrieval, involving (1) multiplex assembly of barcode-tagged constructs, (2) next-generation sequencing (NGS), and (3) selective retrieval of desired constructs by PCR amplification.

Figure 7 is a flowchart illustrating rapid iterative permuting of massively multipart systems using the method of the present disclosure.

Figure 8 is a flowchart illustrating the assembly of cistrons from mixtures of parts (for example, promoters, ribosomal binding sites/ coding regions, and terminators). Here, multiplex scarless stitching is used, but other assembly methods are also compatible with this approach.

Figure 9 is a flowchart illustrating the tagging of a multiplex assembled pool with barcodes and the amplification of the resultant tagged pool.

Figure 10 is a flowchart illustrating the preparation of a tagged pool for sequencing by NGS.

Figure 11 is a diagram showing an assembled nif cistron and listing the set of genetic parts that can be placed at each location within the cistron.

Figures 12A-12D are a series of histograms showing the results of sequencing a library of nif cistrons assembled using presently disclosed methods, including 1.5 million error-free tagged cistrons (Figure 12A) and 55,820 cistrons with unique part combinations (Figure 12B). For each refactor nif CDS, an average of 70% of possible combinations (Figure 12C) and 20% error-free consensus reads (Figure 12D) were found.

Figures 13A-13E are a series of graphs showing that the lower coverage of some pools is largely due to uneven representation of parts, including (A) promoters, (B) ribosomal binding sites (RBS), (C) terminators, (D) 5' scars, and (E) 3' scars.

Figures 14A and 14B are a series of graphs showing (A) frequencies of error types and (B) junction deletions for

promoters and terminators in 7.5 million tagged and sequenced cistrons.

Figures 15A and 15B are flowcharts illustrating methods of tag-directed retrieval of sequence verified DNA molecules, including (A) gene synthesis from pooled oligos and (B) dial-out PCR.

Figure 16 is a diagram showing iterative whole cluster response surface methodology (RSM) design for combinatorial nif operon expression. Figures 18A-18D are a series of graphs showing mutation frequency in PCR retrieval and means for compensating.

Figure 17A-17D. Figures 17A and 17B show mutation frequencies depending on length of construct retrieved (A) or polymerase fidelity (B). Figure 17C shows the number of copies of template cistron retreated after a certain number of PCR cycles. Figure 17D shows the potential sequencing capacity of the MiSeq and HiSeq platforms.

Figure 18 is a diagram describing alternative sequencing strategies for fully sequencing cistrons and larger genetic designs.

Figure 19 is a flowchart illustrating how a barcode-tagged genetic design approach can be used with a high-throughput screen to quantitatively track performance of a large number of permuted designs.

Figure 20 is a flowchart illustrating a streamlined DNA assembly workflow featuring massively one-pot assembly.

Figure 21 is a flowchart illustrating how massively one-pot assembly can be used to reduce the steps in hierarchical cluster assembly into a series of single assembly reactions. In the example shown here, the assembly reactions produce, in turn, 16 scarless nif cistrons, four quarter clusters (1-4, 5-8, 9-12, and 13-16), and one whole refactored cluster (1-16).

Figures 22A-22D are an electrophoresis gel and a series of graphs showing size distribution and extent of reaction (~) for a series of assembly reactions performed as described in Figure 21.

Figure 23 is a diagram illustrating the many possible reactions occurring in a single massively one-pot restriction-ligation reaction, including restriction digest of parts, step-growth polymerization of digested parts, and ligation of competing cohesive ends to growing chains.

Figure 24 is an electrophoresis gel showing that the engineered reaction conditions produce full-length product but nonspecific assembly.

Figure 25 is a series of diagrams illustrating the testing of pairwise assembly to identify design rules for cohesive end specificity.

Figure 26 is a diagram showing refined design rules for cohesive ends and illustrating how the refined design rules were used to computationally redesign sets of orthogonal cohesive ends to improve specificity of assembly 15-fold.

Figure 27 is a pair of electrophoresis gels showing assembly reactions of 16 parts using redesigned scars.

Figure 28 is a graph showing that using the redesigned scars and new reaction conditions resulted in a dramatic increase in the number of colonies containing the correctly sized, full-length insert, from 0% to 77%.

Figure 29 is a diagram illustrating the implementation of unique parts to determine failure modes by sequencing and possible future RSM reaction optimization.

Figure 30 is a set of schematics showing the potential biosynthetic pathways that can be permuted with the disclosed methods.

Figure 31 is a schematic showing how gene expression elements can be tuned by refactoring for precision gene expression.

Figure 32 is a schematic showing how permuted gene clusters can be built from parts making up gene expression elements.

Figure 33 is a schematic showing the optimization of hierarchical gene expression using energy landscapes to define optima in an orthogonal array of genes.

Figure 34 is a schematic showing the construction of gene expression libraries using retrievable gene expression elements.

Figure 35 is a schematic showing the exemplary design of a library of nitrogen fixation gene clusters using the disclosed combinatorial approach.

Figure 36 is a set of graphs and schematics showing the fidelity and errors encountered during assembly of cistrons.

Figure 37 is a schematic showing the iterative process of permuting a multi-gene pathway using the disclosed methods.

Figure 38 is a schematic showing that cistrons retrieved from a pool show very little error in sequence, demonstrating the robust nature of the disclosed methods.

Figure 39 is a schematic showing the assembly of gene expression elements using microfluidics to position the pieces of the assembly in individual wells of a microtiter plate.

Figure 40 is a schematic showing the assembly of a diverse library of nitrogen fixation gene clusters.

Figure 41 is a graph showing the readout, as exemplified by ethylene product of the different nitrogen fixation gene clusters.

Figure 42 is a schematic showing activity as mapped to the cluster designs.

Figure 43 shows that multivariate modeling can be used to optimize the gene cluster design.

Figure 44 is a set of graphs showing multidimensional response surface resolved by DOE.

Figure 45 is a schematic showing how the model based design can be used to predict the response of a refactored gene cluster in an iterative fashion.

Figure 46 is a graph showing the increased activity of the model driven iteratively design optimization of the nitrogen fixation gene cluster.

Figure 47A-47C is a schematic showing a DNA hashing method for multi-module genetic constructs. (Figure 47A) Each modular DNA unit ranging from approximately 100-10,000 bp is assigned a unique DNA hash, which contains a unique identifying DNA sequence. (Figure 47B) Each module can be a composition of discrete DNA parts or shorter DNA modules, such as a combinatorial library of transcriptional units. The DNA hash can include flanking DNA sequences used to PCR amplify a unique internal identifying DNA sequence. (Figure 47C) The genetic composition of a large multi-modular DNA construct can be determined by sequencing the short concatenation of module hashes as opposed to the entire construct. The identifying DNA hash for each module position within a design is amplified and concatenated together into a single DNA fragment. As shown, concatenation for individual designs can be achieved by emulsion overlap extension PCR, wherein each droplet contains ≤ 1 DNA construct (template). In this way, identification of all parts in a large multi-module construct for millions of designs can be achieved using next-generation ("next-gen") sequencing ("NGS").

Figure 48 is a schematic showing the generation of hash tables for synthetic DNA modules. Each hashed module can be a combination of parts, such as a combinatorial library of transcriptional units (as shown). Each unique module or combination of smaller DNA parts is assigned a unique 'hash' DNA barcode, which can be a specified sequence or random DNA sequence determined by sequencing a pool of modules. There can be 10,000's through millions of modules in a library.

Figure 49 is a block diagram depicting a computing machine and a module, in accordance with certain example embodiments.

Figure 50 is schematic showing biological circuit design.

Figure 51 is an electronic image of nucleic acid gel.

Figure 52 is a schematic showing exemplary embodiments of the disclosed method.

DETAILED DESCRIPTION

I. Terms

[0016] Unless otherwise noted, technical terms are used according to conventional usage. Definitions of common terms in molecular biology can be found in Benjamin Lewin, *Genes IX*, published by Jones and Bartlet, 2008 (ISBN 0763752223); Kendrew et al. (eds.), *The Encyclopedia of Molecular Biology*, published by Blackwell Science Ltd., 1994 (ISBN 0632021829); and Robert A. Meyers (ed.), *Molecular Biology and Biotechnology: a Comprehensive Desk Reference*, published by VCH Publishers, Inc., 1995 (ISBN 9780471185710); and other similar references.

[0017] As used herein, the singular forms "a," "an," and "the," refer to both the singular as well as plural, unless the context clearly indicates otherwise. For example, the term "a nucleic acid part" includes single or plural a nucleic acid part and can be considered equivalent to the phrase "at least one a nucleic acid part."

[0018] As used herein, the term "comprises" means "includes." Thus, "a nucleic acid part" means "including a nucleic acid part" without excluding other elements.

[0019] Although many methods and materials similar or equivalent to those described herein can be used, particular suitable methods and materials are described below. In case of conflict, the present specification, including explanations of terms, will control. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

[0020] To facilitate review of the various embodiments of the disclosure, the following explanations of terms are provided:

[0021] Amplification: To increase the number of copies of a nucleic acid molecule. The resulting amplification products are typically called "amplicons." Amplification of a nucleic acid molecule (such as a DNA or RNA molecule) refers to use of a technique that increases the number of copies of a nucleic acid molecule (including fragments).

[0022] An example of amplification is the polymerase chain reaction (PCR), in which a sample is contacted with a pair of oligonucleotide primers under conditions that allow for the hybridization of the primers to a nucleic acid template in the sample. The primers are extended under suitable conditions, dissociated from the template, re-annealed, extended, and dissociated to amplify the number of copies of the nucleic acid. This cycle can be repeated. The product of amplification can be characterized by such techniques as electrophoresis, restriction endonuclease cleavage patterns, oligonucleotide hybridization or ligation, and/or nucleic acid sequencing. An additional example of amplification is the propagation of cells harboring a vector containing an origin of replication, such as a plasmid.

[0023] Other examples of *in vitro* amplification techniques include quantitative real-time PCR; reverse transcriptase PCR (RT-PCR); real-time PCR (rt PCR); real-time reverse transcriptase PCR (rt RT-PCR); nested PCR; strand displace-

ment amplification (see U.S. Patent No. 5,744,311); transcription-free isothermal amplification (see U.S. Patent No. 6,033,881, repair chain reaction amplification (see WO 90/01069); ligase chain reaction amplification (see European patent publication EP-A-320 308); gap filling ligase chain reaction amplification (see U.S. Patent No. 5,427,930); coupled ligase detection and PCR (see U.S. Patent No. 6,027,889); and NASBA™ RNA transcription-free amplification (see U.S. Patent No. 6,025,134) amongst others.

[0024] Antibody: A polypeptide including at least a light chain and/or heavy chain immunoglobulin variable region (or fragment thereof) which specifically recognizes and binds an epitope of an antigen, such as a protein, or a fragment thereof. Antibodies can include one or more of a heavy and a light chain, each of which has a variable region, termed the variable heavy (V_H) region and the variable light (V_L) region. The term also includes recombinant forms such as chimeric antibodies (for example, humanized murine antibodies), heteroconjugate antibodies (such as, bispecific antibodies). An antibody or fragment thereof may be multispecific, for example, bispecific. Antibodies include all known forms of antibodies and other protein scaffolds with antibody-like properties. For example, the antibody can be a monoclonal antibody, a polyclonal antibody, human antibody, a humanized antibody, a bispecific antibody, a monovalent antibody, a chimeric antibody, a single chain antibody (scab), a nanobody, a diabody, an immunoconjugate, or a protein scaffold with antibody-like properties, such as fibronectin or ankyrin repeats. The antibody can have any of the following isotypes: IgG (for example, IgG1, IgG2, IgG3, and IgG4), IgM, IgA (for example, IgA1, IgA2, and IgAsec), IgD, or IgE.

[0025] In most mammals, including humans, whole antibodies have at least two heavy (H) chains and two light (L) chains connected by disulfide bonds. Each heavy chain includes a heavy chain variable region (V_H) and a heavy chain constant region (C_H). However, single chain V_{HH} variants, such as found in camelids, and fragments thereof, are also included. The heavy chain constant region includes three domains, C_H1, C_H2, and C_H3 and a hinge region between C_H1 and C_H2. Each light chain includes a light chain variable region (V_L) and a light chain constant region. The light chain constant region includes the domain, C_L. The V_H and V_L regions can be further subdivided into regions of hyper-variability, termed complementarity determining regions (CDR), interspersed with regions that are more conserved, termed framework regions (FR). Each V_H and V_L is composed of three CDRs and four FRs, arranged from amino-terminus to carboxy-terminus in the following order: FR1, CDR1, FR2, CDR2, FR3, CDR3, FR4. The variable regions of the heavy and light chains contain a binding domain that interacts with an antigen.

[0026] Included are intact immunoglobulins and the variants and binding portions of them well known in the art, such as Fab fragments, Fab' fragments, F(ab')₂ fragments, single chain Fv proteins ("scFv"), and disulfide stabilized Fv proteins ("dsFv") Fd, Feb, or SMIP. An antibody fragment may be, for example, a diabody, triabody, affibody, nanobody, aptamer, domain antibody, linear antibody, single-chain antibody, or multispecific antibodies formed from antibody fragments. Examples of antibody fragments include: (i) a Fab fragment: a monovalent fragment consisting of V_L, V_H, C_L, and C_H1 domains; (ii) a F(ab')₂ fragment: a bivalent fragment including two Fab fragments linked by a disulfide bridge at the hinge region; (iii) a Fd fragment: a fragment consisting of V_H and C_H1 domains; (iv) a Fv fragment: a fragment consisting of the V_L and V_H domains of a single arm of an antibody; (v) a dAb fragment: a fragment including V_H and V_L domains; (vi) a dAb fragment: a fragment consisting of a V_H domain or a V_{HH} domain (such a Nanobody™); (vii) a dAb fragment: a fragment consisting of a V_H or a V_L domain; (viii) an isolated complementarity determining region (CDR); and (ix) a combination of two or more isolated CDRs which may optionally be joined by a synthetic linker. Furthermore, although the two domains of the Fv fragment, V_L and V_H, are coded for by separate genes, they can be joined, using recombinant methods, for example, by a synthetic linker that enables them to be made as a single protein chain in which the V_L and V_H regions pair to form monovalent molecules (known as single chain Fv (scFv)). Antibody fragments may be obtained using conventional techniques known to those of skill in the art, and may, in some instances, be used in the same manner as intact antibodies. Antigen-binding fragments may be produced by recombinant DNA techniques or by enzymatic or chemical cleavage of intact immunoglobulins. An antibody fragment may further include any of the antibody fragments described above with the addition of additional C-terminal amino acids, N-terminal amino acids, or amino acids separating individual fragments.

[0027] An antibody may be referred to as chimeric if it includes one or more variable regions or constant regions derived from a first species and one or more variable regions or constant regions derived from a second species. Chimeric antibodies may be constructed, for example, by genetic engineering. A chimeric antibody may include immunoglobulin gene segments belonging to different species (for example, from a mouse and a human).

[0028] A human antibody refers to a specific binding agent having variable regions in which both the framework and CDR regions are derived from human immunoglobulin sequences. Furthermore, if the antibody contains a constant region, the constant region also is derived from a human immunoglobulin sequence. A human antibody may include amino acid residues not identified in a human immunoglobulin sequence, such as one or more sequence variations, for example, mutations. A variation or additional amino acid may be introduced, for example, by human manipulation. A human antibody of the present disclosure is not chimeric.

[0029] Antibodies may be humanized, meaning that an antibody that includes one or more complementarity determining regions (for example, at least one CDR) substantially derived from a non-human immunoglobulin or antibody is manipulated to include at least one immunoglobulin domain having a variable region that includes a variable framework region

substantially derived from a human immunoglobulin or antibody or includes framework region residues of non-human origin that includes residues that have been replaced by less immunogenic residues more common in comparable human framework regions.

[0030] Genetic part: Any unit of DNA with a proscribed beginning and end and nucleotide sequence.

[0031] Assembly, concatemerization, or DNA assembly: The process of constructing a DNA concatemer by joining one or more DNA parts to each other, according to the methods of the disclosure. Examples of DNA assembly include ligation of DNA parts and ligation-independent methods, such as *in vitro* recombination, overlap extension PCR, round-the-horn PCR mutagenesis, ligation-independent cloning, and homologous recombination. The assembly may be constructed in stepwise fashion by adding one part to a growing concatemer per reaction, or may be assembled in a single reaction using "one-pot" assembly methods as described herein. DNA assemblies constructed according to the methods of the disclosure can be used as parts in further assembly reactions. Exemplary assemblies include higher-order parts, cistrons (for example, cistrons encoding antibodies or other polypeptides), cistron concatemers, and cistron clusters.

[0032] Category of part: A group of parts that share a defined structure and/or function. For example, coding DNA sequences ("CDSs") are a category of parts defined as DNA sequences that encode a polypeptide or a non-coding RNA. A category of parts may be designed such that variants thereof are flanked by one or two characteristic scar sequences, such that they ligate specifically with other pre-defined categories of parts (for example, a different category of parts or the same category of parts). A given DNA part may belong to more than one category of part. For example, a cistron concatemer, defined as any concatemer of more than one cistron, may also be an operon (in which the cistrons are transcribed together in a single transcript) or cistron cluster (in which the complete set of cistrons encodes a set of polypeptides that, together, perform a desired function, for example, biosynthesis pathway components).

[0033] Cistron: a series of genetic elements (for example, one of more of the following: promoters, 5' untranslated regions, ribosomal binding sites, coding regions, introns, exons, 3' untranslated regions, terminators, scars, enhancers, silencers, and/or any combination thereof) present on a single DNA strand. A cistron will generally be a gene containing a plurality of the genetic elements listed above, ordered in such a way that when it is transcribed into RNA, the resultant transcriptional product has a biological function or activity, such as an mRNA that can be translated into a polypeptide (for example, a protein, domain, antibody, or fragment thereof) or splice variants thereof, or a non-coding RNA (ncRNA) that can be processed into, for example, snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs, and long ncRNAs, a ribosomal RNA (rRNA), a transfer RNA (tRNA), or incorporated into a ribozyme or other ribonucleoprotein. An operon can comprise a plurality of genes or cistrons and further comprise one or more parts or elements. The genetic elements in a cistron may be directly contiguous with each other or may be separated by stretches of DNA.

[0034] Concatemer: A polynucleotide construct including two or more pieces of nucleic acid, such as DNA, attached to each other. The pieces of nucleic acid can be arranged, for example, in a pre-specified order or in a random order. The two or more pieces of nucleic acid may be immediately adjacent to each other or separated by, for example, spacers or scars. The two or more pieces of nucleic acid in a concatemer may be, for example, identical parts or belong to the same category of nucleic acid parts. Alternatively, the two or more pieces of nucleic acid may be, for example, different nucleic acid parts or belong to different categories of nucleic acid parts. The two or more pieces of nucleic acid may also belong to different orders of parts, such that one is a higher-order part relative to another. For example, a cistron can be attached to a cistron cluster, a promoter can be attached to a cistron, or a terminator can be attached to a concatemer of cistron components. Alternatively, the two or more nucleic acid pieces may belong to the same order of parts (for example, a promoter attached to an RBS, an RBS attached to a CDS, a CDS attached to a terminator, or a cistron attached to another cistron).

[0035] Covalently linked: Refers to a covalent linkage between atoms by the formation of a covalent bond characterized by the sharing of pairs of electrons between atoms. In one example, a covalent link is a bond between an oxygen and a phosphorous, such as phosphodiester bonds in the backbone of a nucleic acid strand.

[0036] Detectable label: A compound or composition that is conjugated directly or indirectly to another molecule to facilitate detection of that molecule. Specific, non-limiting examples of labels include fluorescent tags, enzymatic linkages, and radioactive isotopes. In some examples, a label is attached to a nucleic acid to facilitate detection of the nucleic acid specifically binds. In specific examples, a detectable label includes a nucleic acid barcode, such as described herein.

[0037] DNA sequencing: The process of determining the nucleotide order of a given DNA molecule. Generally, the sequencing can be performed using automated Sanger sequencing (AB13730xl genome analyzer), pyrosequencing on a solid support (454 sequencing, Roche), sequencing-by-synthesis with reversible terminations (ILLUMINA® Genome Analyzer), sequencing-by-ligation (ABI SOLiD®) or sequencing-by-synthesis with virtual terminators (HELISCOPE®). In some embodiments, the identity of a nucleic acid is determined by DNA or RNA sequencing. Generally, the sequencing can be performed using automated Sanger sequencing (AB13730xl genome analyzer), pyrosequencing on a solid support (454 sequencing, Roche), sequencing-by-synthesis with reversible terminations (ILLUMINA® Genome Analyzer), sequencing-by-ligation (ABI SOLiD®) or sequencing-by-synthesis with virtual terminators (HELISCOPE®); Molecule sequencing (see Voskoboynik et al. eLife 2013 2:e00569 and US Patent Application No. 13/608,778, filed Sep 10, 2012); DNA nanoball sequencing; Single molecule real time (SMRT) sequencing; Nanopore DNA sequencing; Sequencing by

hybridization; Sequencing with mass spectrometry; and Microfluidic Sanger sequencing.

[0038] Hybridization: Oligonucleotides and their analogs hybridize by hydrogen bonding, which includes Watson-Crick, Hoogsteen or reversed Hoogsteen hydrogen bonding, between complementary bases. Generally, nucleic acid consists of nitrogenous bases that are either pyrimidines (cytosine (C), uracil (U), and thymine (T)) or purines (adenine (A) and guanine (G)). These nitrogenous bases form hydrogen bonds between a pyrimidine and a purine, and the bonding of the pyrimidine to the purine is referred to as "base pairing." More specifically, in traditional Watson-Crick pairing, A will hydrogen bond to T or U, and G will bond to C. "Complementary" refers to the base pairing that occurs between two distinct nucleic acid sequences or two distinct regions of the same nucleic acid sequence.

[0039] "Specifically hybridizable" and "specifically complementary" are terms that indicate a sufficient degree of complementarity such that stable and specific binding occurs between the oligonucleotide (or its analog) and the DNA or RNA target. The oligonucleotide or oligonucleotide analog need not be 100% complementary to its target sequence to be specifically hybridizable. An oligonucleotide or analog is specifically hybridizable when there is a sufficient degree of complementarity to avoid non-specific binding of the oligonucleotide or analog to non-target sequences under conditions where specific binding is desired. Such binding is referred to as specific hybridization.

[0040] Isolated: An "isolated" biological component (such a nucleic acid) has been substantially separated or purified away from other biological components in the cell of the organism in which the component naturally occurs, for example, extra-chromatin DNA and RNA, proteins and organelles. The term also embraces nucleic acids prepared by recombinant expression in a host cell as well as chemically synthesized nucleic acids. It is understood that the term "isolated" does not imply that the biological component is free of trace contamination, and can include nucleic acid molecules that are at least 50% isolated, such as at least 75%, 80%, 90%, 95%, 98%, 99%, or even 100% isolated.

[0041] Nucleic acid (molecule or sequence): A deoxyribonucleotide or ribonucleotide polymer including without limitation, cDNA, mRNA, ncRNA, genomic DNA, and synthetic (such as chemically synthesized) DNA or RNA or hybrids thereof. The nucleic acid can be double-stranded (ds) or single-stranded (ss). Where single-stranded, the nucleic acid can be the sense strand or the antisense strand. Nucleic acids can include natural nucleotides (such as A, T/U, C, and G), and can also include analogs of natural nucleotides, such as labeled nucleotides. The major building blocks for polymeric nucleotides of DNA are deoxyadenosine 5'-triphosphate (dATP or A), deoxyguanosine 5'-triphosphate (dGTP or G), deoxycytidine 5'-triphosphate (dCTP or C) and deoxythymidine 5'-triphosphate (dTTP or T). The major building blocks for polymeric nucleotides of RNA are adenosine 5'-triphosphate (ATP or A), guanosine 5'-triphosphate (GTP or G), cytidine 5'-triphosphate (CTP or C) and uridine 5'-triphosphate (UTP or U). Inosine (I) may also be incorporated into synthesized DNA or RNA.

[0042] In some examples, nucleotides include those nucleotides containing modified bases, modified sugar moieties, and modified phosphate backbones, for example as described in U.S. Patent No. 5,866,336 to Nazarenko et al. Examples of modified base moieties which can be used to modify nucleotides at any position on its structure include, but are not limited to: 5-fluorouracil, 5-bromouracil, 5-chlorouracil, 5-iodouracil, hypoxanthine, xanthine, acetylcytosine, 5-(carboxyhydroxymethyl) uracil, 5-carboxymethylaminomethyl-2-thiouridine, 5-carboxymethylaminomethyluracil, dihydrouracil, beta-D-galactosylqueosine, inosine, N⁶-isopentenyladenine, 1-methylguanine, 1-methylinosine, 2,2-dimethylguanine, 2-methyladenine, 2-methylguanine, 3-methylcytosine, 5-methyl cytosine, N⁶-adenine, 7-methylguanine, 5-methylaminomethyluracil, methoxyaminomethyl-2-thiouracil, beta-D-mannosylqueosine, 5'-methoxycarboxymethyluracil, 5-methoxyuracil, 2-methylthio-N⁶-isopentenyladenine, uracil-5-oxyacetic acid, pseudouracil, queosine, 2-thiocytosine, 5-methyl-2-thiouracil, 2-thiouracil, 4-thiouracil, 5-methyluracil, uracil-5-oxyacetic acid methylester, uracil-S-oxyacetic acid, 5-methyl-2-thiouracil, 3-(3-amino-3-N-2-carboxypropyl) uracil, 2,6-diaminopurine and biotinylated analogs, amongst others. Examples of modified sugar moieties which may be used to modify nucleotides at any position on its structure include, but are not limited to arabinose, 2-fluoroarabinose, xylose, and hexose, or a modified component of the phosphate backbone, such as phosphorothioate, a phosphorodithioate, a phosphoramidothioate, a phosphoramidate, a phosphordiamidate, a methylphosphonate, an alkyl phosphotriester, or a formacetal or analog thereof.

[0043] Nucleic acid barcode, barcode, unique molecular identifier, or UMI: A short sequence of nucleotides (for example, DNA, RNA, or combinations thereof) that is used as an identifier for an associated molecule, such as a nucleic acid. A nucleic acid barcode or UMI can have a length of at least, for example, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 60, 70, 80, 90, or 100 nucleotides, and can be in single- or double-stranded form. One or more nucleic acid barcodes and/or UMIs can be attached, or "tagged," to a target molecule and/or target nucleic acid, such as the nucleic acids described herein. This attachment can be direct (for example, covalent or noncovalent binding of the barcode to the target molecule) or indirect (for example, via an additional molecule, for example, a specific binding agent, such as a nucleic acid molecule). Target nucleic acids can be labeled with multiple nucleic acid barcodes in combinatorial fashion, such as a nucleic acid barcode concatemer. DNA sequences can be tagged in this manner at 5' and/or 3' ends. Tagging of DNA sequences with multiple barcodes (for example, one 5' barcode and one 3' barcode) in combinatorial fashion can greatly expand the number of unique identifiers possible within a particular barcode pool. Barcodes can be used as parts in the assembly methods of the present disclosure, or can be produced as scars after Type IIs restriction digest and subsequent ligation.

[0044] Nucleic acid element: A nucleic acid construct or concatemer made from two or more nucleic acid parts, for example an assembly, such as an ordered assembly, of two or more nucleic acid parts.

[0045] Nucleic Acid Part, Nucleic Acid Expression Part, Part, or DNA part: A sequence of nucleic acids, such as RNA or DNA, that can be concatemerized with other nucleic acid parts to form a desired assembly. Exemplary parts include: barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/ cohesive ends, promoters, spacers, ribosomal binding sites (RBS), polypeptide-encoding sequences (coding region or CDS), cis-acting regulatory elements (for example, enhancers, silencers, or insulators), splice sites, 5' untranslated regions (UTRs), 3' UTRs, DNA sequences encoding one or more non-coding RNAs (for example, tRNAs, rRNAs, snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs, or ncRNAs), genes, cistrons, cistron concatemers, cistron clusters, operons, DNA assemblies, or any combination thereof.

[0046] Order: The complexity of the parts used in, for example, an assembly method of the present disclosure. "First-order parts" are parts constructed by traditional DNA cloning or synthesis techniques (for example, de novo synthesis, polymerase chain reaction, restriction digest, or any other common method of producing DNA fragments as known in the art). Examples of first-order parts may include barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/ cohesive ends, promoters or promoter fragments (including but not limited to naturally occurring or genetically engineered promoters, including but not limited to a transcription start site, transcription factor binding site, enhancer element and/or repressor element), spacers, ribosomal binding sites (RBS), polypeptide-encoding sequences (coding region or CDS), cis-acting regulatory elements (for example, enhancers, silencers, or insulators), splice sites, 5' untranslated regions (UTRs), 3' UTRs, DNA sequences encoding one or more non-coding RNAs (ncRNAs), for example, tRNAs, rRNAs, snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs and/or piRNAs. A DNA assembly built from first-order parts using, for example, the methods of the present disclosure is a first order assembly. Exemplary first-order assemblies include CDSs, cistrons, antibody-encoding sequences, or fragments thereof. A part that is itself constructed of, for example, first-order parts is a "second-order part," and is "higher-order" relative to a first-order part. A higher-order part can be constructed from parts of any lower order (for example, a third-order part can be constructed from any combination of first- or second-order parts). Thus, nth-order parts are those constructed from, for example, any number of and/or combination of (n-x)th order parts, in which x can be 1 or greater (for example, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10).

[0047] Primers: Short nucleic acid molecules, such as a DNA oligonucleotide, for example sequences of at least 6 or more nucleotides, or at least 15 nucleotides, wherein one skilled in the art will understand that the primer length depends upon the complexity of the nucleotide target population that it will probe as well as upon hybridization conditions, and which can be annealed to a complementary nucleic acid molecule by nucleic acid hybridization to form a hybrid between the primer and the nucleic acid strand. A primer can be extended along the nucleic acid molecule by a polymerase enzyme. Therefore, primers can be used to amplify a nucleic acid molecule, wherein the sequence of the primer is specific for the nucleic acid molecule, for example so that the primer will hybridize to the nucleic acid molecule under very high stringency hybridization conditions. The specificity of a primer increases with its length. Thus, for example, a primer that includes 30 consecutive nucleotides will anneal to a sequence with a higher specificity than a corresponding primer of only 15 nucleotides. Thus, to obtain greater specificity, probes and primers can be selected that include at least 6, 15, 20, 25, 30, 35, 40, 45, 50 or more consecutive nucleotides.

[0048] In particular examples, a primer is at least 6 nucleotides in length, such as at least 6 contiguous nucleotides complementary to a nucleic acid molecule. Particular lengths of primers that can be used to practice the methods of the present disclosure, include primers having at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least 16, at least 17, at least 18, at least 19, at least 20, at least 21, at least 22, at least 23, at least 24, at least 25, at least 26, at least 27, at least 28, at least 29, at least 30, at least 31, at least 32, at least 33, at least 34, at least 35, at least 36, at least 37, at least 38, at least 39, at least 40, at least 45, at least 50, or more contiguous nucleotides complementary to the target nucleic acid molecule to be amplified, such as a primer of 6-60 nucleotides, 15-50 nucleotides, or 15-30 nucleotides.

[0049] Primer pairs can be used for amplification of a nucleic acid sequence, for example, by PCR, real-time PCR, or other nucleic-acid amplification methods known in the art. An "upstream" or "forward" primer is a primer 5' to a reference point on a nucleic acid sequence. A "downstream" or "reverse" primer is a primer 3' to a reference point on a nucleic acid sequence. In general, at least one forward and one reverse primer are included in an amplification reaction. PCR primer pairs can be derived from a known sequence, for example, by using computer programs intended for that purpose such as Primer (Version 0.5, © 1991, Whitehead Institute for Biomedical Research, Cambridge, MA).

[0050] Methods for preparing and using primers are described in, for example, Sambrook et al. (1989) Molecular Cloning: A Laboratory Manual, Cold Spring Harbor, New York; Ausubel et al. (1987) Current Protocols in Molecular Biology, Greene Publ. Assoc. & Wiley-Intersciences. In one example, a primer includes a label.

[0051] Restriction site: A DNA sequence that includes both the recognition site and cleavage site for a restriction enzyme (for example, a Type IIs restriction enzyme). The cleavage site may be located within the recognition site or may be separated from the recognition site by one or more nucleotides. The cleavage site may be cut at the same position on both strands, thus yielding a "blunt end" with no overhanging single-stranded regions, or may be cut at

different positions on each strand, thus yielding "sticky ends" with overhangs.

[0052] Recognition site: The sequence of nucleotides recognized by a restriction enzyme (for example, a Type IIs restriction enzyme). Recognition sites for restriction enzymes may contain the cleavage site for the enzyme or may be separated from the cleavage site by one or more nucleotides. In either case, the recognition site dictates the location of enzymatic cleavage by the restriction enzyme.

[0053] Retrieval, Retrieve, or Retrieving: refer to the process of isolating one or more desired DNA concatemers from a mixture or pool of DNA concatemers. Exemplary retrieval methods include polymerase chain reaction (PCR), gene synthesis from pooled oligos, or dial-out PCR. The DNA concatemers to be retrieved may be, for example, assembled from DNA parts according to the methods of the disclosure. Generally, the DNA concatemers will be tagged with one or more DNA barcodes (for example, two DNA barcodes), the sequences of which can be used to design primers suitable for PCR retrieval and amplification.

[0054] Sequence identity/similarity: The identity/similarity between two or more nucleic acid sequences, or two or more amino acid sequences, is expressed in terms of the identity or similarity between the sequences. Sequence identity can be measured in terms of percentage identity; the higher the percentage, the more identical the sequences are. Homologs or orthologs of nucleic acid or amino acid sequences possess a relatively high degree of sequence identity/similarity when aligned using standard methods. Methods of alignment of sequences for comparison are well known in the art. Various programs and alignment algorithms are described in: Smith & Waterman, Adv. Appl. Math. 2:482, 1981; Needleman & Wunsch, J. Mol. Biol. 48:443, 1970; Pearson & Lipman, Proc. Natl. Acad. Sci. USA 85:2444, 1988; Higgins & Sharp, Gene, 73:237-44, 1988; Higgins & Sharp, CABIOS 5:151-3, 1989; Corpet et al., Nuc. Acids Res. 16:10881-90, 1988; Huang et al. Computer Appls. in the Biosciences 8, 155-65, 1992; and Pearson et al., Meth. Mol. Bio. 24:307-31, 1994. Altschul et al., J. Mol. Biol. 215:403-10, 1990, presents a detailed consideration of sequence alignment methods and homology calculations. The NCBI Basic Local Alignment Search Tool (BLAST) (Altschul et al., J. Mol. Biol. 215:403-10, 1990) is available from several sources, including the National Center for Biological Information (NCBI, National Library of Medicine, Building 38A, Room 8N805, Bethesda, MD 20894) and on the Internet, for use in connection with the sequence analysis programs blastp, blastn, blastx, tblastn, and tblastx. Blastn is used to compare nucleic acid sequences, while blastp is used to compare amino acid sequences. Additional information can be found at the NCBI web site.

[0055] Once aligned, the number of matches is determined by counting the number of positions where an identical nucleotide or amino acid residue is presented in both sequences. The percent sequence identity is determined by dividing the number of matches either by the length of the sequence set forth in the identified sequence, or by an articulated length (such as 100 consecutive nucleotides or amino acid residues from a sequence set forth in an identified sequence), followed by multiplying the resulting value by 100. For example, a nucleic acid sequence that has 1166 matches when aligned with a test sequence having 1554 nucleotides is 75.0 percent identical to the test sequence ($1166 \div 1554 * 100 = 75.0$). The percent sequence identity value is rounded to the nearest tenth. For example, 75.11, 75.12, 75.13, and 75.14 are rounded down to 75.1, while 75.15, 75.16, 75.17, 75.18, and 75.19 are rounded up to 75.2. The length value will always be an integer. In another example, a target sequence containing a 20-nucleotide region that aligns with 20 consecutive nucleotides from an identified sequence as follows contains a region that shares 75 percent sequence identity to that identified sequence (i.e., $15 \div 20 * 100 = 75$).

[0056] Type IIs restriction enzyme: A restriction enzyme that cuts DNA at a site separated from its recognition site by a defined distance, for example, 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides. Generally, for a Type IIs restriction enzyme, each of the nucleotides present between the recognition site and the cleavage site, and within the cleavage site itself, may be any nucleotide. A Type IIs restriction enzyme may cleave both DNA strands at a position equidistant from the recognition site, leaving a blunt end. Alternatively, a Type IIs restriction enzyme may cleave each DNA strand at a different position relative to the recognition site, leaving a "sticky end" containing a single-stranded overhang sequence, for example, an overhang sequence having a length of 1, 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides. Preferably, the overhang sequence is 4 nucleotides in length, for example, a 4 nucleotide overhang. Blunt ends or sticky ends left after Type IIs restriction digest can be ligated with complementary DNA ends (for example, the other end cleaved off in the digestion reaction, or another end generated by a similar digestion reaction). The sequence of nucleotides present in the overhang sequence can be used for ordered assembly of parts according to the methods of the present disclosure. Exemplary Type IIs restriction enzymes include BbsI, BsaI, and MlyI.

[0057] Variant, Variant of a part, or Variant of parts: is meant members of a category of parts that differ, for example, in sequence, length, or any other aspect of nucleic acid structure. Variants of a given part may be interchangeable in a given location within a DNA assembly. Variants of a given part may also be flanked by one or two characteristic scar sequences, such that a plurality of variants of the part (for example, all variants of the part) ligate specifically with members of other pre-defined categories of parts. Variants of an individual part can have, for example, 70, 80, 90, 95, 99, or more sequence identity.

[0058] Suitable methods and materials for the practice or testing of this disclosure are described below. Such methods and materials are illustrative only and are not intended to be limiting. Other methods and materials similar or equivalent

to those described herein can be used. For example, conventional methods well known in the art to which this disclosure pertains are described in various general and more specific references, including, for example, Sambrook et al., Molecular Cloning: A Laboratory Manual, 2d ed., Cold Spring Harbor Laboratory Press, 1989; Sambrook et al., Molecular Cloning: A Laboratory Manual, 3d ed., Cold Spring Harbor Press, 2001; Ausubel et al., Current Protocols in Molecular Biology, Greene Publishing Associates, 1992 (and Supplements to 2000); Ausubel et al., Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology, 4th ed., Wiley & Sons, 1999; Harlow and Lane, Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, 1990; and Harlow and Lane, Using Antibodies: A Laboratory Manual, Cold Spring Harbor Laboratory Press, 1999. In addition, the materials, methods, and examples are illustrative only and not intended to be limiting.

II. Introduction

[0059] The present disclosure provides methods for engineering complex genetic systems in a realizable time frame. Disclosed herein are methods for streamlined DNA assembly as well as sequence verification and retrieval thereof. Specifically, the disclosure features efficient and high-throughput assembly of many genetic designs from modular "parts," which can be tagged with unique barcodes and sequence verified using, for example, next-generation sequencing technologies. In some embodiments, the sequence is verified prior to assembly. In some embodiments, the sequence is verified after assembly. Desired sequence-verified permutations can be subsequently retrieved by, for example by using tag-directed PCR retrieval techniques. Thus, the present disclosure allows for large libraries of high-fidelity permuted designs to be assembled, verified, tested, and retrieved in short order, permitting rapid and inexpensive iterative search of a genetic design space of interest.

[0060] The present disclosure provides methods to modularize and optimize multi-gene pathways and/or biological systems in iterative fashion, thus increasing the size and complexity of genetic systems that can be engineered. Engineering and optimizing the performance of series of multiple genetic elements and/or large multi-gene systems requires combinatorially tuning a desired biological output (for example, gene expression, gene activity, rate of biosynthesis, biological pathway activity, and reduction in minimal required parts) through many design iterations.

[0061] Engineering genetic systems (for example, multiple genes, pathways, or proteins), can involve chemically ligating a specified set of members, such as designed DNA parts, to construct whole complex designs. Making large sets of combinations with multiple parts in each design can be a laborious and expensive process using traditional techniques. Thus, the number of specified designs that can be constructed is limited. Large, diverse libraries can be constructed in multiplex (for example, all in one solution), but individually testing a specified design in the library requires individually reconstructing that design.

[0062] The new high-throughput DNA assembly strategies disclosed herein leverage next-generation sequencing (NGS) to speed this process and enable statistical combinatorial design for targeted searches through the high-dimensional space of possible genetic designs. By assembling and sequencing designs in multiplex, this approach enables a whole new scale of genetic design (for example, an increase in library size of at least 100-fold to 1000-fold). Thus, the methods disclosed herein can be used, for example, to make large sets of any DNA composites either all combinations from set of parts with designed pairings or many entirely distinct designs with little homology at the assembly junctions using different cloning methods. In certain embodiments, there may be a set of specified permutations for each design, or permutations may be generated randomly. Design permutations can, for example, involve varying parts and/or the sequences of such parts, such as cistron components, cistrons, or concatemers thereof, to produce constructs in combinatorial fashion, which can in turn be experimentally tested. The constructs can, for example, exhibit high fidelity and can be sequence-verified by methods disclosed herein. Once a set of designs have been tested, the results can be used to navigate the genetic design search space and determine whether the experiments, points, and/or designs should be revised or redesigned for iterative improvement in the next design cycle. For example, many permutations of a particular category of parts can be iteratively inserted into a desired framework, sequence-verified, and tested in succession, with the selection of future permutations based on the performance of previously-tested permutations. The methods of the disclosure permit the use of large libraries of permuted designs (for example, at least 500 to 1000 designs per design cycle). This strategy can be applied to any combinatorial search (for example, searching a system of at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, or 16 dimensions).

[0063] The methods described herein provides the capability to efficiently create any assembly from the full combinatorial space described by the product of every potential permutation of 1 through m genetic parts into a set of modules and further assembling every combination of these modules from 1 through n into an assembly as described by the multiplicative product below:

$$Library = \prod_{k=1}^n \prod_{i=1}^m Part_{ik}$$

where Library is the total library size, Part is the individual part type, m is the set of parts in a given module, n is the set of modules.

[0064] Furthermore the method enables the high throughput retrieval of any part in this library via a minimal set of steps where each assembly using traditional pairwise methods would have required

inn-1i

steps, this novel method enables a reduction in steps, all compatible with robotic automation of only n-1 steps.

III. Description of Several Exemplary Embodiments

A. High-Throughput DNA Assembly for Combinatorial Design

[0065] Disclosed are methods for generating a set of nucleic acid elements for the combinatorial construction of a genetic design. Embodiments of the disclosed method include: providing a plurality of nucleic acid parts, comprising one or more categories of nucleic acid parts; assembling the nucleic acid parts to form one or more nucleic acid elements, wherein the nucleic acid elements comprise at least two sequences selected from the plurality of nucleic acid parts; and assembling onto the nucleic acid elements or parts a 3' and/or a 5' flanking nucleic acid sequence that orders assembly of higher-order nucleic acid constructs, facilitates retrieval and/or identification of the nucleic acid element. In certain embodiments, the disclosed methods involve providing a plurality of nucleic acid parts that include one or more of: a set of nucleic acid sequences comprising different promoters; a set of nucleic acid sequences comprising a set of different ribosomal binding sites; a set of nucleic acid sequences encoding a polypeptide; and/or a set of nucleic acid sequences comprising a set of different transcription terminators.

[0066] In some embodiments, a plurality of nucleic acid parts includes one or more of mammalian, insectoid, fungal, or bacterial expression elements. In one embodiment, a plurality of nucleic acid parts includes one or more of 5' UTR processing ribozyme (insulator) parts, natural repressors, phage polymerases, natural activators, sigma factors, programmable repressors, programmable activators, orthogonal ribosomes, antisense sRNA repressors, activating riboregulators, sRNA regulators converted to affect transcription, antisense RNA transcriptional attenuators, dCas9 repression, and dCas9 activation barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/cohesive ends, promoters, spacers, ribosomal binding sites (RBSs), polypeptide-encoding sequences (coding region or CDS), cis-acting regulatory elements (for example, enhancers, silencers, or insulators), splice sites, 5' untranslated regions (UTRs), 3' UTRs, DNA sequences encoding one or more non-coding RNAs (for example, tRNAs, rRNAs, snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, piRNAs, or ncRNAs), genes, cistrons, cistron concatemers, cistron clusters, operons, DNA assemblies and those described in Nielsen et al., Current Opinion in Chemical Biology 2013, 17:878-892.

[0067] The nucleic acid parts are assembled, for example using the exemplary techniques described herein, to form one or more nucleic acid elements. These nucleic acid elements include at least two sequences selected from the plurality of nucleic acid parts. Subsequently, the sequence of the sequence of the nucleic acid elements are determined, for example by sequencing (such as NGS), amplification, hybridization or any combination thereof, thereby providing a set of sequence-verified nucleic acid elements. In certain embodiments, a nucleic acid barcode is assembled onto the 5' and/or 3' end of the nucleic acid element, for example to facilitate retrieval of the nucleic acid element from a pool of such elements, such as retrieval by amplification (for example PCR) and/or nucleic acid isolation. As disclosed herein, the assembly can be sequential or concurrent. Exemplary methods of each are given below. The assembly is typically via ligation, such as with a ligase, although chemical joining of the nucleic acids is contemplated (for example chemical joining followed by amplification such the amplification products appear identical to the amplification products of ligated nucleic acids). Methods of chemical joining and ligation are well known in the art. In some embodiments, the assembly comprises Golden Gate Assembly and/or assembly using Type IIs restriction enzymes, Gibson Assembly, restriction ligation, ligation cycling reaction, ligation independent cloning, Gateway cloning, homologous recombination, and/or iterative capped assembly.

[0068] A variety of methods can be used for nucleic acid assembly according to the methods of the present disclosure. In certain embodiments, scarless stitching methodologies are used to assemble concatemers of nucleic acids, such as DNA, parts. Scarless stitching allows for the ligation of two DNA sequences without the production of a polynucleotide "scar," an intervening sequence having, for example, a defined sequence and length (for example, a 4 nucleotide scar). As scar sequences can interfere with the activity and/or expression of the assembled design, scarless stitching is well-suited for use in assembly according to the methods of the present disclosure. Sequential rounds of scarless stitching can be used to add additional parts, such as nucleic acid parts to a growing concatemer, or to ligate two concatemers. For example, sequential rounds of scarless stitching can be used to assemble concatemers of at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 40, 50, 75, or 100 parts, such as nucleic acid parts, for example nucleic acid parts. Scarless stitching methods known in the art can be used to directionally order the assembly of parts (for example, assembling parts AB and CD only in the order ABCD, and not in the orders ABDC, BACD, or DCAB). Alternatively, scarless stitching can be un-ordered (for example, ordered randomly).

In certain embodiments, the parts used to construct concatemers that are of the same order. In other embodiments, the parts used to construct concatemers that are of different order (for example, some parts are higher-level relative to other parts).

[0069] One method of cloning that is useful in the disclosed method is the Golden Gate cloning method devised by Engler et al. (PLoS ONE 3(11): e3647, 2008; Golden Gate cloning utilizes Type IIs restriction enzymes, which cleave DNA outside of the confines of their recognition sites. For example, BsaI binds to double-stranded DNA at the recognition sequence GGTCTC, but cleaves the DNA molecule at a four base pair (bp) region separated from the recognition sequence by one nucleotide. Thus, the complete site of action for BsaI has the sequence GGTCTC plus NWXYZ, wherein NWXYZ can be any set of five contiguous nucleotides. The enzymatic cleavage product includes one section of DNA containing the BsaI recognition sequence and the reverse complement of WXYZ left as an overhang, and a second section of DNA containing the WXYZ overhang. Notably, this second section of DNA no longer contains a BsaI recognition site. One can design a construct in which a part, such as a nucleic acid part, for example a nucleic acid part is flanked by two restriction sites, such as Type IIs restriction sites, such that cleavage will result in the removal of the recognition sites from both ends of the resultant fragment. The cohesive ends containing the Type IIs recognition site are cleaved off in this scenario. If the nucleic acid part is ligated with a similarly-constructed fragment based on overhang complementarity, the enzyme recognition site will be eliminated and the construct will no longer be cleavable by the enzyme. Thus, constructs can be designed such that only a desired ligation (for example, ligation between two parts or ligation between a part and a desired backbone or vector) results in elimination of the enzyme recognition site on both sides of the ligation junction. Undesired ligations (for example, ligation between a nucleic acid part and a cohesive end or ligation between two cohesive ends) will recreate the entirety of the Type IIs restriction site, including the recognition site, and are thus reversible reactions. Thus, the reaction can be funneled towards the desired product. Because the overhang sequences can be made of any combination of four nucleotides, the ligation junction can have any sequence desired. For example, using the first or last four nucleotides of the 3' or 5' fragment, respectively, can be used to produce a scarless junction. Alternatively, overhang sequences can be designed to intentionally introduce a 4-nucleotide scar, for example, a scar having a predefined sequence. Additional methods for assembly that can be used according to the methods of the present disclosure include, for example, BioBrick™ assembly, BglBrick assembly, overlap extension polymerase chain reaction (OE-PCR), Gibson isothermal assembly, transformation-associated recombination, pairwise selection, InFusion™, sequence-and ligation-independent cloning (SLIC), uracil-specific excision reagent (USER) cloning, Bacillus Domino recombination, (Ellis et al., Integr. Biol. 3: 109-118, 10 2011), Golden Braid cloning (Sarrion-Perdigones et al., PLoS ONE 6(7): e21622, 2011), and ligation based methods or homologous recombination-based methods known in the art (Merryman and Gibson, Metabol. Eng. 14: 196-204, 2012).

[0070] In certain embodiments, the present disclosure provides a refinement of the Golden Gate cloning method, in which multiple parts (for example, at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 40, 50, 75, 100, 150, 200, 250, or 256 nucleic acid parts) are digested and ligated into concatemers in a single reaction ("one-pot" assembly, illustrated in Figure 23). The 4-bp overhang scar sequences that are produced after Type IIs restriction digest can be used as codes to direct a desired ligation between two nucleic acid parts, one having a particular 4-bp scar overhang, and the other having the reverse complement overhang. Placing Type IIs restriction sites on both 5' and 3' ends of a given part, such as a nucleic acid part allow it to be assembled to two other parts simultaneously. In certain embodiments, the one-pot reaction mixture contains: a plurality of nucleic acid, each with one or two Type IIs restriction sites flanking the nucleic acid part; one or more Type IIs restriction enzymes capable of cleaving the Type IIs restriction sites, and a ligase suitable for attaching nucleic acid parts after overhang annealing. The reaction may involve restriction digest of nucleic acid parts to produce parts containing overhangs and complementary cohesive end fragments, step-growth polymerization of digested nucleic acid parts, and competing cohesive end ligation to growing chains of nucleic acid parts. Importantly, only the concatemerization/polymerization of two nucleic acid parts, a nucleic acid part and a growing chain, or two growing chains, are irreversible due to the loss of the Type IIs recognition site. In some embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators includes at least one Type IIs restriction site and, prior to assembly, they are cleaved at the Type IIs restriction site to produce sticky ends to facilitate assembly. In some embodiments, cleavage of the Type IIs restriction site results in cohesive end overhangs at the 5' and 3' ends of the nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators such that the sequences are assembled in the corrected read order.

[0071] The extent of reaction can be increased by, for example, increasing reaction time, increasing the ratio of Type IIs enzyme to ligase, increasing the number of temperature cycles, removing cohesive ends, modulating the duration of

the cycles, or modulating the order of part assembly. Reaction yield can be raised by, for example, increasing DNA concentration or increasing enzyme concentration. Golden Gate overhang sequences can further be used to specify the order in which multiple nucleic acid parts are assembled in a one-pot assembly reaction (for example, assembling parts AB, CD, and EF in the order ABCDEF). This is accomplished by employing multiple unique 4-bp overhang scar sequences. Each unique overhang sequence will only ligate with another overhang sequence of suitable sequence complementarity. Because restriction digest and ligation freely occur in the same reaction in one-pot assembly, the multiple parts can be cleaved (to produce overhangs), annealed together, and ligated in a desired order, all within a single one-pot reaction. Thus, n distinct overhang sequences can be used to simultaneously assemble $n+1$ distinct DNA fragments in a specified order, as demonstrated in Example 2 (see Figure 21). For example, one-pot assembly can be used to assemble concatemers of a pre-specified order of nucleic acid parts, including at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 40, 50, 75, 100, 150, 200, 250, or 256 parts. In certain embodiments, the nucleic acid parts used to construct a concatemer are of the same order. In other embodiments, the nucleic acid parts used to construct a concatemer are of different order (for example, some nucleic acid parts are higher-level relative to other nucleic acid parts). The present disclosure further features optimized design rules for 4-bp scar sequences and improved reaction conditions that permit consistent assembly of full-length products in an ordered fashion. Exemplary design rules and reaction conditions are described in Examples 2 and 3.

[0072] In some embodiments, the nucleic acid elements include at least three sequences selected from the set of a promoters, ribosomal binding sites, coding sequences, and transcription terminators. In some embodiments, the nucleic acid elements include at least four sequences selected from the set of a promoters, ribosomal binding sites, coding sequences, and transcription terminators. In specific embodiments two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same set. In some embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are from the same set. In certain embodiments, the nucleic acid element includes a functional cistron or a subunit thereof. In some embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators are sequence variants.

[0073] Nucleic acid parts that can be used in the assembly methods of the present disclosure can include any sequence of DNA, ranging in size and scale from individual components of a cistron to clusters of cistrons. Parts can range in size from less than 50 bp to greater than 8 kilobases (kb). Exemplary nucleic acid parts may have a length of about 10 bp, 20 bp, 30 bp, 40 bp, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 9 kb, or 10 kb, or a length in a range specified by any combination of the prior-listed lengths (for example, ranges between adjacent listed lengths).

[0074] A cistron requires certain components to be suitable for expressing a gene product in vitro or in vivo. Non-limiting examples of cistron components that can be used as nucleic acid parts in the present disclosure include one or more of any combination of the following: barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/cohesive ends, promoters, spacers, ribosomal binding sites (RBS), and/or sequences encoding a polypeptide of interest (coding region or CDS), for example, an entire polypeptide of interest or a subset (for example, a domain). Figure 11 depicts an exemplary cistron containing, in order from 5' to 3' end: a 5' barcode sequence, a BbsI recognition site, a scar sequence, a promoter/spacer, an RBS, a CDS, a terminator/spacer, a second scar sequence, a second BbsI recognition site, and a 3' barcode sequence. Such nucleic acid parts may be initially designed by, for example, mining a database such as NCBI to identify known parts for each desired category of nucleic acid parts, such as those listed above. The components of a cistron can vary depending on the system or organism in which it will be expressed. For example, the RBS can be a 5' cap, a Shine-Dalgarno sequence, or an internal ribosomal entry site (IRES). The CDS can include exons and/or introns, although generally it will contain coding sequences.

[0075] In some embodiments, the CDS encodes a protein (for example, an antibody). The antibody can be, for example, in any of the following forms: IgG, IgM, IgA, IgD, IgE, Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP and/or as defined above. The antibody can be, for example, a mouse, rat, rabbit, goat, bovine, canine, chicken, donkey, feline, guinea pig, hamster, horse, monkey, sheep, or pig antibody. In certain embodiments, the antibody is, for example, a human antibody, a chimeric antibody, or a humanized antibody. Exemplary antibody components or antibody fragments that can be used as parts in the methods of the present disclosure include Fab fragments, Fc fragments, F(ab')₂ fragments, Fab' fragments, Fv fragments, IgG fragments, variable domains, constant domains, heavy chains, light chains, framework sequences, antigen-binding sites, and complementarity-determining regions.

[0076] In certain embodiments, the CDS encodes a fragment of a protein, for example, a protein domain, antibody component, antibody fragment, or polypeptide. In certain embodiments, at least one of the parts encodes an antibody variable domain. In certain embodiments, at least one of the parts encodes an antibody constant domain. In certain embodiments, at least one of the nucleic acid parts encodes a protein from a biological pathway. In certain embodiments, the set of parts encodes multiple proteins from a biological pathway. In certain embodiments, the set of parts encodes

all of the proteins from a biological pathway. In certain embodiments, the biological pathway includes a metabolic pathway. In certain embodiments, the metabolic pathway includes a synthetic pathway or a catabolic pathway. In certain embodiments, the biological pathway includes a genetic pathway. In certain embodiments, the biological pathway includes a signal transduction pathway; for example, synthetic genetic circuits.

[0077] Additional components that can be used as parts in methods of the present disclosure include, for example, cis-acting regulatory elements (for example, enhancers, silencers, or insulators), splice sites, 5' untranslated regions (UTRs), 3' UTRs, and/or sequences encoding one or more non-coding RNAs (ncRNAs) for example tRNAs, rRNAs, snoRNAs, microRNAs, siRNAs, snRNAs, exRNAs, and/or piRNAs. 5' UTR processing ribozyme (insulator) parts, natural repressors, phage polymerases, natural activators, sigma factors, programmable repressors, programmable activators, orthogonal ribosomes, antisense sRNA repressors, activating riboregulators, sRNA regulators converted to affect transcription, antisense RNA transcriptional attenuators, dCas9 repression, and dCas9 activation barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/ cohesive ends, promoters, spacers, ribosomal binding sites (RBS), polypeptide-encoding sequences (coding region or CDS), genes, cistrons, cistron concatemers, cistron clusters, operons, DNA assemblies and those described in Nielsen et al., *Current Opinion in Chemical Biology* 2013, 17:878-892.

[0078] Parts may contain epigenetic modifications (for example, DNA methylation), non-DNA nucleotides (for example, uracil), and/or nucleotide analogues. In certain embodiments, parts assembled by methods of the present disclosure can contain a single, for example, promoter, but multiple RBSs and/or encoding regions. Assembly of parts may be performed in sequential fashion, in which one nucleic acid part is added to a growing concatemer at a time, or by one-pot assembly. Alternatively, concatemers of parts (for example, a promoter and RBS) may be assembled to form a subassembly and then subsequently attached to additional parts and/or subassemblies. Parts may be assembled in a specified order or in an un-ordered (for example, randomly ordered) fashion.

[0079] In some embodiments, the methods further include amplifying the nucleic acid elements to form a mixture of amplified sets of nucleic acid elements, for example amplifying the nucleic acid elements by polymerase chain reaction (PCR). In some embodiments, the nucleic acid elements are inserted into nucleic acid vectors, for example to generate a plurality of plasmids containing the nucleic acid elements. These nucleic acid vectors can also be amplified, for example by propagation in a suitable host cell, for example under selective pressure. Examples of plasmid propagation are well known in the art and include propagation in bacterial cells such as but not limited to *E. coli*. Isolation of such plasmids is equally known in the art. In some embodiments, the plurality of plasmids are pooled.

[0080] Particular sequence-verified nucleic acid elements assembled according to methods of the present disclosure can be readily retrieved from a pool of assembled nucleic acid elements, for example, using PCR. Thus, in certain embodiments, the disclosed methods include retrieving one or more of the sequence-verified, for example by nucleic acid amplification and/or isolation. For example, barcoded assembled nucleic acid elements can be retrieved from, for example from a pool, by using a unique set of primers that specifically hybridize the barcodes on the 5' and 3' ends of the bar coded nucleic acid elements. In some examples, multiple assembled nucleic acid elements are retrieved using primers that specifically hybridize to the barcodes. Because each assembly is flanked by a pair of unique barcodes, a given assembly may be retrieved by performing PCR using primers targeting the barcodes. Additional methods for tag-directed retrieval of sequence-verified assemblies include gene synthesis from pooled oligos (Figure 15A; Kosuri et al., *Nat. Biotechnol.* 28(12): 1295-1299, 2010; US Patent Publication No. US 20140045728 A1) and dial-out PCR (Figure 15B; Schwartz et al., *Nat. Meth.* 9: 913-915, 2012; US Patent Publication No. US 20120283110 A1).

[0081] Aspects of the present disclosure also concern the construction of higher-order nucleic acid constructs. Higher-order nucleic acid constructs are assemblies built from a plurality of nucleic acid elements, for example, using methods of the present disclosure. Thus, disclosed herein are methods of assembling higher-order nucleic acid constructs. The methods include retrieving two or more nucleic acid elements from the sequence-verified pool of nucleic acid elements, assembling the two or more nucleic acid elements to form one or more higher-order nucleic acid constructs, and determining the sequence of the higher-order nucleic acid constructs thereby generating a sequence-verified pool of higher-order nucleic acid constructs. Higher-order nucleic acid construct can include concatemers of multiple cistron components, cistrons, or cistron concatemers, and can include concatemers of multiple operon components, operons, operon concatemers, and can include concatemers of concatemers. One higher-order nucleic acid construct may be higher in order relative to a second higher-order nucleic acid construct if the first higher-order nucleic acid construct is, for example, an assembly of parts belonging to the same category of nucleic acid parts as the second part or an assembly of such assemblies of parts. Such higher-order nucleic acid construct can also be used in the assembly methods described herein to construct higher-order assemblies. In certain embodiments, a higher-order nucleic acid construct includes two or more nucleic acid sequences encoding a protein. In specific examples, the protein includes an antibody or an antibody fragment, such as a humanized antibody or antibody fragment. In some examples the antibody or antibody fragment is of isotype IgG, IgA, IgM, IgE, or IgD, and/or the fragment selected includes a Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP or as provided above in the definition of an antibody.

[0082] Exemplary higher-order assemblies include cistrons, cistron concatemers, and cistron clusters as well as operons, operon concatemers and operon clusters. In certain embodiments, the higher-order assembly is an antibody (for

example, a chimeric antibody or humanized antibody) or antibody fragment.

[0083] Higher-order nucleic acid constructs can be ligated with other higher-order nucleic acid construct or with lower-order parts such as those described above. Assembly of higher-order nucleic acid construct into higher-order assemblies may be performed sequentially, for example, by adding one higher-order nucleic acid construct onto a growing concatemer per reaction. Alternatively, multiple higher-order nucleic acid construct (for example, more than three higher-order nucleic acid constructs) may be assembled into a concatemer in a single reaction. In a further embodiment, parts can be added to a higher-order assembly in iterative fashion.

[0084] In certain embodiments, a plurality of cistrons can be assembled into a cistron concatemer. A cistron concatemer includes at least two cistrons (for example, at least 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, or 32 cistrons), and may contain all of the cistrons in a biological pathway (for example, a metabolic pathway, such as a synthetic pathway or a catabolic pathway, genetic pathway, or a signal transduction pathway, such as those described herein). The cistrons in a cistron concatemer assembled using methods of the present disclosure can be arranged in a specified order, for example, by using a unique scar sequence for each cistron junction, such that a given cistron can only ligate with one specific cistron at each end. Alternatively, the cistrons in a cistron concatemer can be arranged in an un-ordered (for example, randomly ordered) fashion. In certain embodiments, the higher-order nucleic acid construct includes two or more nucleic acid sequences encoding two or more proteins from a biological pathway.

[0085] In certain embodiments, cistron concatemers are used as higher-order nucleic acid constructs for further assembly. For example, as shown in Figure 21, groups of cistrons can be assembled into first-order cistron concatemers by Golden Gate cloning. The first-order cistrons can be amplified, purified, excised from their plasmid backbones, and further assembled by Golden Gate cloning into second-order "quarter clusters," which are in turn ultimately assembled by Golden Gate cloning into a third-order whole cluster. It is contemplated that whole clusters can be used as higher-order nucleic acid constructs in further assembly reactions, as well.

[0086] In certain embodiments, a nucleic acid barcode is assembled to the 5' and/or 3' end of the higher-order nucleic acid construct. In some embodiments, the higher-order nucleic acid constructs are amplified, for example by PCR to form a mixture of amplified, higher-order nucleic acid constructs. In certain embodiments, the higher-order nucleic acid constructs are inserted into a nucleic acid vectors, thereby generating a plurality of plasmids containing the higher-order nucleic acid constructs.

[0087] Nucleic acid barcodes are frequently used to tag DNA, RNA, or other molecules in an identifiable way. Barcodes can be any nucleic acid sequence that may be used to identify a particular tagged molecule. Barcoding methods have been used extensively to allow parallel, multiplexing experimentation of different types of genetic alterations (see, for example, Craig et al., Nat. Methods 5: 887-893, 2008; Gerrits et al., Blood 115: 2610-2618, 2010; and Berns et al., Nature 428: 431-437, 2004). In the methods of the present disclosure, DNA barcodes are used to uniquely tag genetic designs. In some embodiments, two barcodes are attached to a DNA assembly, one at the 5' end and one at the 3' end. The use of two barcodes on the same molecule greatly increases the number of unique identifiers that can be generated using a particular barcode pool. Barcodes that can be used according to the methods of the disclosure can have a length of at least, for example, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 35, 40, 45, 50, 60, 70, 80, 90, or 100 bp. In some embodiments, the barcodes are chip-synthesized and reused, such that the retrieval primers can be reused across multiple libraries. Barcodes can be added to an assembled pool of genetic designs by, for example, multiplex Gibson assembly. In one embodiment, plasmids containing assembled designs are purified and emulsion multiplex PCR is performed to add desired cohesive ends and flanking adaptors to each assembly. The resultant pool of assemblies is ligated to backbone vectors containing a pair of N20 barcodes and an ampicillin resistance (AmpR) cassette, such that the barcodes flank a single assembly after ligation, with each assembly receiving a unique combination of barcodes. In certain embodiments, the barcodes can be used to retrieve a desired individual assembly from a pool. For example, two flanking barcode sequences can be used to design primers for PCR retrieval of the assembly. The primer can include a variant region corresponding to the barcodes and an invariant region that is, for example, common to the assembled parts.

[0088] DNA sequences (for example, constructs assembled according to the methods of the disclosure) can be amplified, for example, at various steps in the methods of the disclosure (for example, after assembly, after barcode tagging, prior to sequencing adaptor ligation, after sequencing adaptor ligation, or after retrieval). Methods for DNA amplification are well-understood in the art. For example, DNA sequences can be amplified by polymerase chain reaction (PCR). PCR primers can be designed based on, for example, barcode or scar sequences. PCR may be preceded by a restriction digest step that excises the target DNA from a vector. Alternatively, DNA can be amplified by, for example, transformation of cells (for example, bacteria or yeast) with, for example, a self-replicating vector (for example, a plasmid, bacterial artificial chromosome, yeast artificial chromosome, fosmid, or virus) containing the target DNA, followed by clonal expansion of the cells in a suitable growth medium. The amplified vectors can be readily purified from the cells using standard preparation techniques in the art.

[0089] The present disclosure provides methods for assembling genetic designs such that the resultant assemblies

can be readily sequence verified, for example, in multiplex. In some embodiments, genetic designs assembled according to the methods of the disclosure are flanked by a pair of barcode sequences. The assembly and barcodes are, in turn, flanked by restriction sites (for example, NotI sites), thus allowing the excision of assembly and flanking barcode pair by a restriction digest. For assemblies short enough in length that sequencing from both ends would permit resolution of the entire assembly sequence (for example, assembly sequences shorter than about 0.5-1 kb), next-generation sequencing (NGS) adaptors are attached and NGS is performed. Particular sequence-verified genetic designs assembled according to methods of the present disclosure can be readily retrieved from a pool of assembled constructs, for example, by hierarchical retrieval. For example, the assembly pool can be tested using a desired assay, such as, for example, gene analysis (for example, RT-PCR, Northern blotting, Western blotting, ELISA, microarray analysis, RNA-Seq, flow cytometry, or proteomic analysis), protein activity analysis (for example, cytochrome P450 activity assays, reporter assays, enzyme-based assays, kinase assays, or phosphatase assays), measurement of biosynthetic product formation or purity, or high-throughput variants of any of the above assays.

[0090] Further disclosed is a method of optimizing a combination of nucleic acid elements for the production of an optimum desired phenotype, the method including: retrieving a plurality of sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs; assembling the retrieved sequences to construct a biological pathway assembly; testing each of the assembled biological pathway assemblies for the desired phenotype; determining which characteristic or characteristics of the assembled biological pathway contributes to the desired phenotype by comparing multiple assemblies; based on the determination, plurality of sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs; assembling the retrieved sequences to construct a biological pathway assembly; testing each of the of the assembled biological pathway assemblies for the desired phenotype; and repeating until a of the assembled biological pathway assemblies for the desired phenotype is identified that produces said optimum desired phenotype.

[0091] Assemblies that perform strongly in the assay may be useful to retrieve from the pool. Because each assembly is flanked by a pair of unique barcodes, a given assembly may be retrieved by performing PCR using primers targeting the barcodes. Additional methods for tag-directed retrieval of sequence-verified assemblies include gene synthesis from pooled oligos (Figure 15A; Kosuri et al., Nat. Biotechnol. 28(12): 1295-1299, 2010; US Patent Publication No. US 20140045728 A1) and dial-out PCR (Figure 15B; Schwartz et al., Nat. Meth. 9: 913-915, 2012; US Patent Publication No. US 20120283110 A1).

[0092] For longer assemblies, there are several possible sequencing strategies. In certain embodiments, one or more "common" parts are identical in all assemblies, such as that schematized in Figure 10. In such cases, the assembly fragment can be self-circularized and then PCR amplified using primers directed at the ends of the common part. The primers can be designed so that the PCR reaction goes outward, away from the center of the common part, such that the resultant PCR product includes the two barcodes at the center, flanked by the parts of the assembly, with the common part removed (except the portions complementary to the PCR primers). This smaller fragment can then be ligated to sequencing adaptors and NGS sequencing can be performed. These strategies should be ideal for sequence verification of cistron-sized assemblies. In other embodiments, the assemblies can be too large for the strategies described above, or the assembly pool can contain no common parts. Under such circumstances, sequencing can be performed, for example, using multiplex "jumping" libraries, primer walking amplification, or Single-Molecule Long Reads (SMRT) approaches (Figure 17).

[0093] Exemplary NGS platforms suitable for use with the present disclosure include the MiSeq and HiSeq platforms (Illumina), the SOLiD platform (Applied Biosystems), Roche 454 pyrosequencing, Ion Torrent, Ion Proton (Life Technologies), PacBio RS, Oxford Nanopore, and other NGS platforms known in the art.

[0094] Aspects of the method further include hashing the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs, for example as described below and in the Figures. Thus in certain embodiments, the method further includes attaching a DNA hash, that include a unique nucleic acid barcode, to each of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs at the 5' and/or 3' end of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs and assigning the individual DNA hashes to the individual nucleic acid sequences in the plurality of nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs, thereby creating a hash table, wherein the sequence of the individual unique nucleic acid barcode in the hash table represents the sequence of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs. In some embodiments, the DNA hash includes nucleic acid primer sites 5' and 3' of a unique molecular identifier. In certain embodiments, the method further includes amplifying the unique molecular identification sequence using primers that specifically hybridize to the nucleic acid primer sites. In some examples the primers further include a nucleic acid sequence that specifically hybridizes to the sequence of one of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs. In some embodiments, a concatemer of amplified unique molecular identification sequences is assembled, and optionally the sequence of the assembled concatemer is determined, thereby determining the sequence of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs and/or the order of assembly of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs.

B. Nucleic acid elements

[0095] Disclosed is a set of sequence-verified nucleic acid elements, or amplification products thereof, for the combinatorial construction of genetic elements including one or more nucleic acid elements assembled from: a set of nucleic acid sequences comprising different promoters; a set of nucleic acid sequences comprising a set of different ribosomal binding sites; a set of nucleic acid sequences encoding a polypeptide; and/or a set of nucleic acid sequences comprising a set of different transcription terminators, wherein sequence of the nucleic acid element is determined after assembly. The nucleic acid parts can include any sequence of DNA, ranging in size and scale from individual components of a cistron to clusters of cistrons. Parts can range in size from less than 50 bp to greater than 8 kilobases (kb). Exemplary nucleic acid parts may have a length of about 10 bp, 20 bp, 30 bp, 40 bp, 50 bp, 60 bp, 70 bp, 80 bp, 90 bp, 100 bp, 200 bp, 300 bp, 400 bp, 500 bp, 600 bp, 700 bp, 800 bp, 900 bp, 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 9 kb, or 10 kb, or a length in a range specified by any combination of the prior-listed lengths (for example, ranges between adjacent listed lengths). Additional nucleic acid parts that can be components of nucleic acid elements include one or more of any combination of the following: barcodes, restriction sites (for example, Type IIs restriction enzyme sites), scars/cohesive ends, promoters, spacers, ribosomal binding sites (RBS), and/or sequences encoding a polypeptide of interest (coding region or CDS), for example, an entire polypeptide of interest or a subset (for example, a domain). Figure 11 depicts an exemplary cistron containing, in order from 5' to 3' end: a 5' barcode sequence, a BbsI recognition site, a scar sequence, a promoter/spacer, an RBS, a CDS, a terminator/spacer, a second scar sequence, a second BbsI recognition site, and a 3' barcode sequence. Such nucleic acid parts may be initially designed by, for example, mining a database such as NCBI to identify known parts for each desired category of nucleic acid parts, such as those listed above. In some embodiments, the nucleic acid parts further comprise one or more nucleic acid assembly sequences that order assembly. In some embodiments, the nucleic acid assembly sequence comprises one or more restriction sites and, prior to assembly, they are cleaved restriction site to produce sticky ends to facilitate assembly.

[0096] In some embodiments, the nucleic acid elements include at least three sequences selected from the set of a promoters, ribosomal binding sites, coding sequences, and transcription terminators. In some embodiments, the nucleic acid elements include at least four sequences selected from the set of promoters, ribosomal binding sites, coding sequences and transcription terminators. In some embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same set. In some embodiments, two or more of the assembled nucleic acid parts in the nucleic acid element are from the same set. In some embodiments, the nucleic acid element includes a functional cistron or a subunit thereof. In some embodiments, the nucleic acid element includes a nucleic acid encoding a protein. In some embodiment the nucleic acid elements or parts include a 3' and/or a 5' flanking nucleic acid sequence that orders assembly of higher-order nucleic acid constructs, facilitates retrieval and/or identification of the nucleic acid element. In some embodiments, the sequence-verified nucleic acid elements further include a nucleic acid barcode to the 5' and/or 3' end of the nucleic acid element. In some embodiments, the 3' and/or 5' flanking sequences comprise one or more restriction sites and, prior to assembly, they are cleaved restriction site to produce sticky ends to facilitate assembly. In some embodiments, the one or more restriction sites comprises at least one Type IIs restriction site. In some embodiments, cleavage of the Type IIs restriction site results in cohesive end overhangs at the 5' and 3' ends of the nucleic acid sequences such that the sequences are assembled into a continuous DNA fragment with the designed order and orientation of its components. In some embodiments, the nucleic acid parts further comprise one or more nucleic acid barcodes. In some embodiments, the 3' and/or 5' flanking sequences comprise one or more nucleic acid barcodes.

[0097] In some embodiments, assembling includes ligation of the nucleic acid sequences. In some embodiments, one or more of the set of nucleic acid sequences comprising the set of different promoters; the set of nucleic acid sequences comprising the set of different ribosomal binding sites; the set of nucleic acid sequences encoding the set of different gene products, such as polypeptides, for example proteins or fragments thereof; and the set of nucleic acid sequences comprising the set of different transcription terminators are sequence variants, for example homologs, orthologs and/or sequence variants by virtue of the redundancy of the genetic code. By way of example, in the construction of a metabolic pathway (or other pathway) one may want to vary the sequences of the component, such as promoter sequences, coding sequences etc., to tune the and or tune the final product. In some embodiments, the protein includes an antibody or an antibody fragment. In some embodiments, the antibody or antibody fragment includes a humanized antibody or antibody fragment. In some embodiments, the antibody or antibody fragment is of isotype IgG, IgA, IgM, IgE, or IgD, and/or the fragment selected includes a Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP. In some embodiments, at least one of the nucleic acid parts encodes an antibody variable domain. In some embodiments, at least one of the nucleic acid parts encodes an antibody constant domain. In some embodiments, at least one of the nucleic acid parts encodes a protein from a biological pathway. In some embodiments, the set of parts encodes multiple proteins from a biological pathway. In some embodiments, the set of parts encodes all of the proteins from a biological pathway. In some embodiments, the biological pathway includes a metabolic pathway. In some embodiments, the metabolic pathway includes a synthetic pathway or a catabolic pathway. In some embodiments, the biological pathway includes a genetic pathway. In some embodiments, the biological pathway includes a signal transduction pathway.

[0098] Also disclosed is a pool of higher-order nucleic acid constructs or amplification products thereof, including one or more nucleic acid elements assembled from the sequence-verified nucleic acid elements. In some embodiments, the higher-order nucleic acid constructs further include a nucleic acid barcode to the 5' and/or 3' end. Disclosed is set of plasmids including the pool of higher-order nucleic acid constructs and/or the set of sequence-verified nucleic acid elements.

C. Kits

[0099] In certain embodiments, the disclosure features a kit providing a pool of sequence-verified nucleic acid elements and/or the pool of higher-order nucleic acid constructs, and a plurality of primers for retrieving one or more sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs. The pool can be assembled, for example, by the methods of the present disclosure. The kit can also include instructions describing which barcodes within the pool are associated with particular constructs. The kit can further include materials for retrieving certain constructs from the pool (for example, oligonucleotide primers for PCR retrieval, a polymerase, and a PCR buffer).

D. Applications

[0100] The present disclosure provides means for assembling complex genetic designs from component parts, and can thus be useful for genetic engineering and synthetic biology applications. Exemplary applications include construction and/or optimization of any of the following: gene clusters, proteins (for example, antibodies; see above), multi-gene networks, genetic circuits, biosynthetic pathways, signal transduction pathways, metabolic pathways, and gene expression regulatory pathways. For example, the genes required for biosynthesis of a useful compound (for example, a drug) can be assembled into a cluster and permutations of each category of parts systematically tested for their effect on attributes of interest (for example, efficiency of drug synthesis, drug potency, drug effectiveness, and level of gene expression in host cells).

[0101] The present disclosure also enables rapid and inexpensive construction and optimization of genetic designs, such as a biological pathway or genetic circuit. The components of such genetic designs (for example, parts as described herein) can be permuted iteratively at a massive scale in order to optimize a genetic design by permuting attributes such as gene expression, genetic architecture, operon organization, synthetic regulation, or transcript splicing. Additionally, a CDS can be permuted and optimized by testing varying CDS sequences to examine the effect of mutations or different synthetic CDS (for example, different polypeptides, codon usage, multiple nucleotide polymorphisms, single nucleotide polymorphisms, deletions, duplications, insertions, and frame shifts). The methods described herein further enable rapid and targeted combinatorial permuting of a genetic design to probe high-dimensional and vast genetic search spaces. Thus, the methods herein can be used to engineer and debug larger and more complex genetic designs than existing methods in the art, thus expanding the complexity of biological systems that can be written, engineered, and optimized.

[0102] Existing methods for assembling genetic permutations produce a variety of error modes relating to, for example, part purity, low sequencing coverage, partial clusters, pipetting error, deletions, and insertions. To determine the prevalence of such error modes, the inventors have generated permutations of gene clusters including components of nitrogenase biosynthesis and nitrogen fixation pathways (Figure 2). Golden Gate hierarchical cloning according to the MoClo system described by Weber et al. (PLoS ONE 6(2): e16765, 2011) was used to generate a library of nif clusters featuring 5 permutations of nifHDKY, 6 permutations of nifENJB, 4 permutations of nifQF, and 1 permutation of nifUS-VWZM (Figure 3). We found that only 65% of assembled sequences showed fully correct clusters, with backbone insertions as the most prevalent error mode, followed by error due to manual pipetting. Further examination of a set of 77 correct clusters revealed an average of 0.17 single nucleotide polymorphisms (SNPs) and 0.47 small deletions within each cluster.

[0103] In addition to a high error rate, these existing methods for DNA assembly are notably costly and slow. As shown in Figure 4, producing each unique nif cluster required 53 iterations of the following procedure: (1) assembly reaction, (2) bacterial transformation, (3) colony selection, (4) plasmid purification, and (5) concentration normalization. Each arrow represents a single iteration of the procedure. In addition, 64 Sanger sequencing reactions were required to obtain the complete sequence of each cluster. Thus, there is a need for high-throughput DNA assembly allowing for rapid and inexpensive production, sequence verification, testing, and retrieval of genetic permutations.

E. DNA Hashing

[0104] Aspects of the current disclosure relate to DNA hashing, either in the context of the other embodiments, or as stand-alone technology. By way of example and with reference to the methods disclosed herein, the size and scale of multi-part designs that can be built in synthetic biology has been rapidly increasing, yet massively parallel testing of combinatorial designs has remained challenged by the ability to track the performance of individual genetic designs in

high-throughput (pooled) screens or selections. Previous approaches used have been: individually constructing and assaying genetic designs (low through-put), combinatorial libraries without tracking all or full designs (sequence individual 'winners'), and combinatorial libraries limited to short regions of DNA variation with high-throughput NGS tracking (for example <5 parts as opposed to 100+ parts as disclosed herein). For large multi-part DNA designs, a key bottleneck to applying high-throughput screening or selections has been the need to sequence the entire (10-100,000 kb) design to determine the composition of parts.

[0105] To solve this problem, the inventors have developed a DNA hashing technology, similar to hashing in a computer environment, where small data sequences, typically transformed with a hashing function, are used to pinpoint or index larger data elements, such as in a database. With reference to the methods disclosed herein, as an analogy to the computer database if one considers the different nucleic acid parts as values in a database, a unique DNA hash, for example barcode can be assigned to the different nucleic acid parts, strings of genetic parts, or entire cistrons such that the individual nucleic acid part, string of parts, or cistron in a larger construct can be determined, both for order and presence, by determining the sequence of the DNA hashes, for example determining the sequence of a DNA hash concatemer. The sequence information, for example the sequences of the DNA hashes and/or nucleic acid sequence elements, can be stored in a hash table. Nested DNA hashing can also be used. For example an individual genetic part or construct of genetic parts can be uniquely barcoded within a cistron construct and the assembled cistron containing additional parts can be uniquely barcoded to represent all parts within the cistron.

[0106] With reference to Figure 47 a unique identifying DNA sequence (e.g. hash that includes a unique DNA barcode) is attached to longer modules of DNA (see Figure 47A) such that only the short barcode 'hash' needs to be sequenced to determine the presence a longer DNA fragment or string of DNA parts (see Figure 47B). After assaying performance for approximately 100,000 to 1 million design variants (e.g., binning or fitness determination), these short 'hash' barcodes are then concatenated into a single DNA string, for example using emulsion PCR stitching (see Figure 47C). In this way, the concatenated hashes of the DNA parts or cistrons have been compressed to potentially the length of a single paired-end read, which allows for increased assaying through-put and lower cost sequencing.

[0107] Disclosed is a method of nucleic acid hashing, for example of a plurality of nucleic acid sequence elements. The method includes attaching a DNA hash, that has a unique nucleic acid barcode to each of the plurality of nucleic acid sequence elements at the 5' and/or 3' end of the nucleic acid sequence; assigning the individual DNA hashes to the individual nucleic acid sequences in the plurality of nucleic acid sequence elements, thereby creating a hash table wherein the sequence of the individual unique nucleic acid barcode in the hash table represents the sequence of the nucleic acid sequence elements. In certain embodiments, the unique DNA hash is assigned randomly, and the unique DNA hash sequence, and/or the individual nucleic acid with which it is associated, and which the DNA hash sequence thereby denotes, is determined by sequencing. In certain embodiments the unique nucleic acid barcode includes flanking nucleic acid primer sites 5' and 3' of a unique molecular identifier, such as a unique 12 bp DNA sequence. In certain embodiments, the method further includes assembling two or more of the plurality of nucleic acid sequences. In certain embodiments, the method further includes amplifying the unique molecular identification sequence using primers that specifically hybridize to the nucleic acid primer sites. In certain embodiments, the primers further comprise a nucleic acid sequence that specifically hybridizes to the sequence of one of the nucleic acid sequence elements. In certain embodiments the method further includes assembling a concatemer of amplified unique molecular identification sequences. In certain embodiments, the method further includes determining the sequence of the assembled concatemer, thereby determining the sequence of the nucleic acid sequence elements and/or the order of assembly of the nucleic acid sequence elements.

[0108] In computing, a hash table (hash map) is a data structure used to implement an associative array, a structure that can map keys to values. A hash table uses a hash function to compute an index into an array of buckets or slots, from which the correct value can be found.

F. Other Embodiments

[0109] Embodiments of the disclosed methods may be implemented in a computer program that embodies the functions described and illustrated herein, wherein the computer program is implemented in a computer system that comprises instructions stored in a machine-readable medium and a processor that executes the instructions. However, it should be apparent that there could be many different ways of implementing embodiments in computer programming, and the embodiments should not be construed as limited to any one set of computer program instructions. Further, a skilled programmer would be able to write such a computer program to implement an embodiment of the disclosed embodiments based on the figures and associated description in the application text. Therefore, disclosure of a particular set of program code instructions is not considered necessary for an adequate understanding of how to make and use embodiments. Further, those skilled in the art will appreciate that one or more aspects of embodiments described herein may be performed by hardware, software, or a combination thereof, as may be embodied in one or more computing systems. Moreover, any reference to an act being performed by a computer should not be construed as being performed by a

single computer as more than one computer may perform the act. Further disclosed are one or more electronic files including sequence information for a pool of sequence-verified nucleic acid elements and/or a pool of higher-order nucleic acid constructs; and a plurality of primers for retrieving one or more sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs.

[0110] The example embodiments described herein can be used with computer hardware and software that perform the methods and processing functions described previously. The systems, methods, and procedures described herein can be embodied in a programmable computer, computer-executable software, or digital circuitry. The software can be stored on computer-readable media. For example, computer-readable media can include a floppy disk, RAM, ROM, hard disk, removable media, flash memory, memory stick, optical media, magneto-optical media, CD-ROM, etc. Digital circuitry can include integrated circuits, gate arrays, building block logic, field programmable gate arrays (FPGA), etc.

[0111] The example systems, methods, and acts described in the embodiments presented previously are illustrative, and, in alternative embodiments, certain acts can be performed in a different order, in parallel with one another, omitted entirely, and/or combined between different example embodiments, and/or certain additional acts can be performed, without departing from the scope and spirit of various embodiments. Accordingly, such alternative embodiments are included in the examples described herein.

[0112] Figure 49 depicts a computing machine 2000 and a module 2050 in accordance with certain example embodiments. The computing machine 2000 may correspond to any of the various computers, servers, mobile devices, embedded systems, or computing systems presented herein. The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 in performing the various methods and processing functions presented herein. The computing machine 2000 may include various internal or attached components such as a processor 2010, system bus 2020, system memory 2030, storage media 2040, input/output interface 2060, and a network interface 2070 for communicating with a network 2080.

[0113] The computing machine 2000 may be implemented as a conventional computer system, an embedded controller, a laptop, a server, a mobile device, a smartphone, a set-top box, a kiosk, a vehicular information system, one or more processors associated with a television, a customized machine, any other hardware platform, or any combination or multiplicity thereof. The computing machine 2000 may be a distributed system configured to function using multiple computing machines interconnected via a data network or bus system.

[0114] The processor 2010 may be configured to execute code or instructions to perform the operations and functionality described herein, manage request flow and address mappings, and to perform calculations and generate commands. The processor 2010 may be configured to monitor and control the operation of the components in the computing machine 2000. The processor 2010 may be a general purpose processor, a processor core, a multiprocessor, a reconfigurable processor, a microcontroller, a digital signal processor ("DSP"), an application specific integrated circuit ("ASIC"), a graphics processing unit ("GPU"), a field programmable gate array ("FPGA"), a programmable logic device ("PLD"), a controller, a state machine, gated logic, discrete hardware components, any other processing unit, or any combination or multiplicity thereof. The processor 2010 may be a single processing unit, multiple processing units, a single processing core, multiple processing cores, special purpose processing cores, co-processors, or any combination thereof. According to certain example embodiments, the processor 2010 along with other components of the computing machine 2000 may be a virtualized computing machine executing within one or more other computing machines.

[0115] The system memory 2030 may include non-volatile memories such as read-only memory ("ROM"), programmable read-only memory ("PROM"), erasable programmable read-only memory ("EPROM"), flash memory, or any other device capable of storing program instructions or data with or without applied power. The system memory 2030 may also include volatile memories such as random access memory ("RAM"), static random access memory ("SRAM"), dynamic random access memory ("DRAM"), and synchronous dynamic random access memory ("SDRAM"). Other types of RAM also may be used to implement the system memory 2030. The system memory 2030 may be implemented using a single memory module or multiple memory modules. While the system memory 2030 is depicted as being part of the computing machine 2000, one skilled in the art will recognize that the system memory 2030 may be separate from the computing machine 2000 without departing from the scope of the subject technology. It should also be appreciated that the system memory 2030 may include, or operate in conjunction with, a non-volatile storage device such as the storage media 2040.

[0116] The storage media 2040 may include a hard disk, a floppy disk, a compact disc read only memory ("CD-ROM"), a digital versatile disc ("DVD"), a Blu-ray disc, a magnetic tape, a flash memory, any other non-volatile memory device, a solid state drive ("SSD"), any magnetic storage device, any optical storage device, any electrical storage device, any semiconductor storage device, any physical-based storage device, any other data storage device, or any combination or multiplicity thereof. The storage media 2040 may store one or more operating systems, application programs and program modules such as module 2050, data, or any other information. The storage media 2040 may be part of, or connected to, the computing machine 2000. The storage media 2040 may also be part of one or more other computing machines that are in communication with the computing machine 2000 such as servers, database servers, cloud storage, network attached storage, and so forth.

[0117] The module 2050 may comprise one or more hardware or software elements configured to facilitate the computing machine 2000 with performing the various methods and processing functions presented herein. The module 2050 may include one or more sequences of instructions stored as software or firmware in association with the system memory 2030, the storage media 2040, or both. The storage media 2040 may therefore represent examples of machine or computer readable media on which instructions or code may be stored for execution by the processor 2010. Machine or computer readable media may generally refer to any medium or media used to provide instructions to the processor 2010. Such machine or computer readable media associated with the module 2050 may comprise a computer software product. It should be appreciated that a computer software product comprising the module 2050 may also be associated with one or more processes or methods for delivering the module 2050 to the computing machine 2000 via the network 2080, any signal-bearing medium, or any other communication or delivery technology. The module 2050 may also comprise hardware circuits or information for configuring hardware circuits such as microcode or configuration information for an FPGA or other PLD.

[0118] The input/output ("I/O") interface 2060 may be configured to couple to one or more external devices, to receive data from the one or more external devices, and to send data to the one or more external devices. Such external devices along with the various internal devices may also be known as peripheral devices. The I/O interface 2060 may include both electrical and physical connections for operably coupling the various peripheral devices to the computing machine 2000 or the processor 2010. The I/O interface 2060 may be configured to communicate data, addresses, and control signals between the peripheral devices, the computing machine 2000, or the processor 2010. The I/O interface 2060 may be configured to implement any standard interface, such as small computer system interface ("SCSI"), serial-attached SCSI ("SAS"), fiber channel, peripheral component interconnect ("PCI"), PCI express ("PCIe"), serial bus, parallel bus, advanced technology attached ("ATA"), serial ATA ("SATA"), universal serial bus ("USB"), Thunderbolt, FireWire, various video buses, and the like. The I/O interface 2060 may be configured to implement only one interface or bus technology. Alternatively, the I/O interface 2060 may be configured to implement multiple interfaces or bus technologies. The I/O interface 2060 may be configured as part of, all of, or to operate in conjunction with, the system bus 2020. The I/O interface 2060 may include one or more buffers for buffering transmissions between one or more external devices, internal devices, the computing machine 2000, or the processor 2010.

[0119] The I/O interface 2060 may couple the computing machine 2000 to various input devices including mice, touchscreens, scanners, electronic digitizers, sensors, receivers, touchpads, trackballs, cameras, microphones, keyboards, any other pointing devices, or any combinations thereof. The I/O interface 2060 may couple the computing machine 2000 to various output devices including video displays, speakers, printers, projectors, tactile feedback devices, automation control, robotic components, actuators, motors, fans, solenoids, valves, pumps, transmitters, signal emitters, lights, and so forth.

[0120] The computing machine 2000 may operate in a networked environment using logical connections through the network interface 2070 to one or more other systems or computing machines across the network 2080. The network 2080 may include wide area networks (WAN), local area networks (LAN), intranets, the Internet, wireless access networks, wired networks, mobile networks, telephone networks, optical networks, or combinations thereof. The network 2080 may be packet switched, circuit switched, of any topology, and may use any communication protocol. Communication links within the network 2080 may involve various digital or an analog communication media such as fiber optic cables, free-space optics, waveguides, electrical conductors, wireless links, antennas, radio-frequency communications, and so forth.

[0121] The processor 2010 may be connected to the other elements of the computing machine 2000 or the various peripherals discussed herein through the system bus 2020. It should be appreciated that the system bus 2020 may be within the processor 2010, outside the processor 2010, or both. According to some embodiments, any of the processor 2010, the other elements of the computing machine 2000, or the various peripherals discussed herein may be integrated into a single device such as a system on chip ("SOC"), system on package ("SOP"), or ASIC device.

[0122] Although specific embodiments have been described above in detail, the description is merely for purposes of illustration. It should be appreciated, therefore, that many aspects described above are not intended as required or essential elements unless explicitly stated otherwise. Modifications of, and equivalent components or acts corresponding to, the disclosed aspects of the example embodiments, in addition to those described above, can be made by a person of ordinary skill in the art, having the benefit of the present disclosure, without departing from the spirit and scope of embodiments defined in the following claims, the scope of which is to be accorded the broadest interpretation so as to encompass such modifications and equivalent structures.

EXAMPLES

[0123] The following examples are provided by way of illustration only and not by way of limitation. Those of skill in the art will readily recognize a variety of noncritical parameters that could be changed or modified to yield essentially the same or similar results.

Example 1

Multiplex scarless cistron assembly and sequencing followed by tag-directed retrieval

[0124] High-throughput DNA assembly and genetic design strategies were used to optimize a pathway of 16 genes. This 16-gene nitrogen fixation (*nif* gene cluster is entirely modular and synthetic ("refactored") with ~100 parts. The synthetic *nif* pathway encodes 16 essential genes that have tightly coordinated action to convert atmospheric nitrogen into ammonia, a critical process for global agriculture productivity. Using our novel assembly strategy, we constructed and sequenced a library of >55,000 unique *nif* transcriptional units (TUs). This is a 100 fold to 1,000-fold improvement in library size over previous studies. Each TU is composed of characterized genetic parts (5-7 each), including one refactored *nif* gene. These *nif* TU modules can be directly assembled into whole 16-gene, 100-part permuted cluster designs (~1045 unique cluster designs possible from this library). Additionally, approximately 7.5 million *nif* cistrons were sequenced to characterize DNA assembly error modes and quantify their frequencies. Next, from the library of *nif* modules systematically permuted refactored clusters were built and tested to investigate optimizing nitrogen fixation activity by tuning expression within the factored cluster using characterized biological parts. Combinatorial design and statistical design of experiments were applied to constrain the part substitutions integrated into the cluster designs fabricated and tested in each round. Since the total number of possible combinations was in the 1040 range, random sampling of the pool was unlikely to produce an optimal design. By determining which parameters have the greatest effect on system performance, we could thus traverse the genetic design space more efficiently. Together, these methods enable model-guided redesign and optimization of genetic system performance with minimized DNA assembly required (~10-fold reduction in cloning).

[0125] Figure 5 illustrates the workflow used for one embodiment of the present disclosure, in which a pool of barcode-tagged constructs was assembled in multiplex. As shown in Figure 6, each construct included a complete *nif* cistron, including four parts that could be added in order: a promoter, RBS, CDS, and terminator. In the present experiment, all cistrons featured the same *nif* CDS, which was included in a part featuring the varying RBS (Figure 7). A unique pair of DNA barcodes flanked each cistron constructed, and next-generation sequencing was used to determine the sequence of each cistron and its associated barcodes. Cistrons were then assembled into clusters and experimentally tested. Desired cistrons and/or clusters were then retrieved by designing barcode-specific primers and performing polymerase chain reaction (PCR). After each round of assembly and testing, design of experiments methodology and learning algorithms were used to traverse the design space for optimal redesign of the next round of cistrons and/or clusters.

[0126] By selecting which parts to include in the assembly reactions, we could restrict the expression search space from the outset (Figure 7). The promoter and RBS/CDS parts were assembled first, followed by addition of terminators, thus producing a pool of thousands of cistrons, which could be tested for genetic design performance. The assembly procedure is illustrated in Figure 8. In brief, equimolar amounts of a "Part 1 Mix" containing promoters and a "Part 2 Mix" containing the variable RBS and constant CDS parts were attached by scarless stitching into intermediate assembly plasmids, which were then amplified by cloning and purified from plated bacterial cells. Scarless stitching was performed by Golden Gate cloning (Engler et al., PLoS ONE 3(11): e3647, 2008). The intermediate assemblies were then mixed with an equimolar amount of a "Part 3 Mix" containing terminator parts to assemble complete cistrons, again by Golden Gate cloning-based scarless stitching. After additional amplification and purification steps, cohesive ends and flanking adapters were added to each cistron to produce a cistron pool. Assuming equal probability of all parts and assemblies, we would expect that approximately ten times the number of unique combinations would be required for a 99.99% probability of sampling all possible permutations.

[0127] Once the pool was assembled, each cistron was uniquely tagged with a pair of barcodes (Figure 9). Each cistron was inserted into a backbone containing an ampicillin resistance cassette and N20 barcodes (including 1E12 and 5E23 barcode pairs) by multiplex Gibson assembly, thus producing circularized plasmid constructs each containing a single cistron flanked by a unique pair of barcodes and NotI restriction sites. Bacterial cells were transformed with the tagged plasmid pool, plated, and grown according to routine microbial culture techniques, thus clonally amplifying individual plasmid constructs.

[0128] By plating a certain number of cells per transformation, this randomly bottlenecked the tagged pool, limiting the diversity of molecules to sequence. Amplified constructs were then recovered by plasmid isolation techniques routine in the art.

[0129] As shown in Figure 10, plasmid DNA containing multiple copies of each tagged cistron was then isolated and digested with NotI. Designs smaller than 500-1000 bp could progress immediately to sequencing adaptor ligation. For designs greater in size than about 500 bp, NotI-digested cistrons were self-circularized using standard ligation protocols performed at dilute concentrations, and then PCR amplified using primers directed at the ends of the common CDS sequence to produce ~500 bp PCR products that include, in order: the terminator, 3' barcode, 5' barcode, promoter, and RBS. Sequencing adaptors were subsequently ligated to both ends of each fragment. The pool was then sequenced using the MiSeq next-generation sequencing platform (Illumina), producing two reads per fragment or cistron, each read

starting from one end of the polynucleotide and progressing inward. For short designs for which self-circularization and PCR amplification were skipped, the resultant contigs include the entirety of the cistron assembly. For longer designs, the contigs reveal the sequence of the 500 bp PCR product. In other case, the two barcode sequences and the sequences of all variable parts (for example, all parts other than the CDS) can be determined, thus allowing for comparison of a particular assembly's performance with its unique structure.

[0130] The structure of an exemplary *nif* cistron is shown in Figure 11. From 5' to 3' end, each cistron included: a 5' barcode sequence, a BbsI recognition site, a 4-bp scar (selected from 10 possible scars), a promoter/spacer (selected from 5 possible promoter/spacer combinations), an RBS (selected from two possible RBSs), a *nif* CDS (selected from 16 possible refactored *nif* CDSs), a terminator/spacer (selected from 5 possible terminator/spacer combinations), a second 4-bp scar (selected from the same set of 10 possible scars as the first scar), a second BbsI recognition site, and a 3' barcode sequence. Using the set of parts described here, 80,000 unique *nif* cistrons can be produced ~5,000 cistrons per *nif* CDS. Each 4-bp scar is located within the overhang region of the adjacent BbsI recognition site to form a complete BbsI restriction site. Thus, BbsI digestion produces overhangs on each end of the cistron that can be used for ordered assembly with other cistrons based on scar sequence complementarity to generate clusters of cistrons.

[0131] Using this method, a library of error-free, unique, sequence-verified *nif* cistrons was constructed, including 2 gigabases (Gb) of error-free cistrons and 72 megabases (Mb) of cistrons having unique part combinations (Figure 12). In total, 1.5 million error-free, tagged cistrons were produced (Figure 12A), of which 55,820 featured unique part combinations (Figure 12B). An average of 70% of all possible combinations was recovered for each of the 16 refactored *nif* CDS pools (Figure 12C). An average of 20% of cistron consensus reads were error-free per refactored *nif* CDS (Figure 13D). The lower coverage observed in some pools was largely due to an uneven representation of parts (Figure 13A-13E). For example, the t0 terminator sequence predominated in each of the refactored *nif* pools (Figure 13C). 7.5 million tagged cistrons were sequenced to identify scarless stitching error modes and frequencies (Figure 14). We found that 3' junction deletion was the most prevalent error mode (Figure 14A), which was corroborated by data showing that the frequency of junction deletions was higher in terminators than in promoters (Figure 14B).

[0132] Next tag-directed retrieval of sequence-verified DNA molecules was performed using two different methods: gene synthesis from pooled oligos (Figure 15A; Kosuri et al., Nat. Biotechnol. 28(12): 1295-1299, 2010) and dial-out PCR (Figure 15B; Schwartz et al., Nat. Meth. 9: 913-915, 2012). Each of the above-listed references is incorporated herein in its entirety. Cistrons from all 167 pools were successfully retrieved, cloned, and sequenced, yielding a 93% error-free sequence rate. This confirms that this approach can be successfully utilized for iterative whole cluster response surface methodology (RSM) design (Figure 16). Of the 44 cistrons sequenced, three contained one to two single nucleotide polymorphisms (SNPs), of which at least one was generated during retrieval. The impact of mutation frequency for PCR retrieval was subsequently found to increase with length of cistron retrieved (Figure 17A) and decrease with improved polymerase fidelity (Figure 17B). The frequency of a template cistron to retrieve is inversely proportional to the number of PCR cycles necessary to yield 20 nM product concentration (Figure 17C). In addition, we found that the Illumina HiSeq platform offers significantly improved potential sequencing capacity over the MiSeq platform (Figure 17D). Alternative approaches for fully sequencing cistrons and larger genetic designs include multiplex "jumping" libraries, primer walking amplification, and Single Molecule Long Reads (Figure 18).

[0133] In summary, the barcoded genetic design approach disclosed herein can be used in combination with a high-throughput screen to quantitatively track the performance of millions of permuted designs more quickly and inexpensively than existing methods in the art (Figure 19). This method is ideal for assembling designs of anywhere from less than 2 kb up to 8 kb, on the order of 1E5 to 1E6 permutations at a time. Assembled genetic permutations can then be tested via high-throughput screening, indeed, the constructs of the disclosure can be used to directly quantify performance by next-generation sequencing to identify and count the number of barcodes present in, for example, high-performing pools versus low-performing pools. High-performing designs can then be selectively retrieved and verified.

Example 2

Streamlining DNA assembly using massively one-pot assembly

[0134] Assembly of designs can be further optimized by minimizing the number of steps required to create ordered concatemers of parts (for example, barcodes, scars, promoters, spacers, RBSs, CDSs, terminators, restriction sites, cistrons, or concatemers of any of the above). A one-pot assembly method was developed for creating ordered cistron concatemers (Figure 20), involving restriction digest and ligation of parts within the same reaction. This method takes advantage of the fact that the 4-bp cleavage sites present in the restriction sites for Type IIs restriction enzymes (for example, BsaI, BbsI, and MlyI) may include any combination of four nucleotides, as the enzyme recognition site is separated from the cleavage site. Thus, we can use the cleavage site as a means for storing information about the identity of the part or cistron as well as establish an ordering system, by using predefined 4-bp sequences for each junction between parts or cistrons to be ordered. This approach allowed for the hierarchical assembly of entire cistron

clusters based on sequence complementarity of the overhangs produced after enzymatic cleavage (Figure 21). For example, cistrons 1,2, and 3 can be assembled in that order if: cistron 1 is flanked by scars "a" and "b," cistron 2 is flanked by scars "b" and "c," and cistron 3 is flanked by scars "c" and "d." This can be used to produce "first-order" concatemers in plasmids, which can in turn be assembled with, for example, other first-order concatemers, clusters, or scarless cistrons, to form larger assemblies (for example, a quarter cluster in a second-order plasmid). This process can be repeated multiple times to produce complete refactored clusters.

[0135] We used two different assembly methods to attempt to construct 16-cistron nif clusters: (1) MoClo, an assembly technique described by Weber et al. (PLoS ONE 6(2): e16765, 2011 and (2) the one-pot hierarchical assembly method (schematized in Figure 23). As shown in Figure 23, many reactions occur simultaneously in each massively one-pot restriction-ligation reaction, including restriction digest of parts (in this case, pre-assembled cistrons) from their cohesive ends, step-growth polymerization of digested parts, and competing cohesive ends ligating to growing chains. Crucially, only polymerization of parts is irreversible, as this reaction results in removal of the restriction enzyme recognition sites from the junction. Upon completion of assembly, residual cohesive ends can be removed by a cleanup step. The resultant clusters produced by each method were quantified for size distribution and the extent of the reaction (ξ) by capillary electrophoresis. An Agilent Bioanalyzer was used to quantify dilute assembly reactions and resolve peaks. Under MoClo conditions, only assembly of up to eight parts was detected (Figure 22A). No 16-part cistron clusters were produced by the MoClo method. In contrast, our one-pot hierarchical assembly method successfully produced full-length 16-cistron nif clusters. Analysis of the assemblies produced by one-pot hierarchical assembly showed that full-length product was consistently produced, but that the assembly was often not specific (Figure 24). We thus endeavored to identify design rules for cohesive end specificity that would permit efficient and consistent ordered assembly. We tested pairwise assembly of parts to determine their assembly specificity (Figure 25), and used the results to develop refined design rules for cohesive ends. In particular, we found that preferred cohesive ends (1) cannot share greater than 2 nucleotides (nt) with any scar, (2) cannot share greater than 2 nt with the reverse complement of any scar or with the cohesive end itself, (3) must contain at least 1 guanine (G) or cytosine (C), (4) cannot contain more than 3 Gs or Cs, and (5) must be exactly four nucleotides in length (Figure 26). We built random sets of cohesive ends that satisfy the refined rules, including up to 24 scars using the original set of 10 scar sequences (and a maximum set size of 32 scars), and performed triplicate runs of 1 million steps (1 day each). As a result, we found a significant reduction in non-orthogonality (Figure 26) compared to cohesive ends produced using the original design rules. We then performed assembly reactions of 16 parts with the redesigned scars and new reaction conditions (Figure 27). 77% of the resultant colonies contained the full-length insert, compared to 0% from assemblies built using the original scars (Figure 28). In summary, disclosed herein a method for using massively one-pot Golden Gate cloning for streamlined DNA assembly. We have engineered conditions to increase extent of reaction and yield, and have tested specificity with 70 pairwise assemblies, which we used to computationally design scar sets following refined design rules. As a result, we have improved scar specificity 15-fold, such that for 16-part assemblies, 77% of colonies contain full-length clusters. Implementation of unique parts in the one-pot assembly method described herein us being used to determine failure modes by sequencing, and for possible future RSM reaction optimization (Figure 29). In particular, whole nif clusters are being assembled containing 16 unique cistrons (21 kb length) and a maximal scar set with short parts containing 31 unique parts (1 kb length). Combinatorial optimization of the reaction conditions is also being performed to further improve yields and extent of reaction. Lastly, we are testing differentiable parts for assembly of larger parts and parts in greater numbers.

Example 3.

Engineered One-Pot Golden Gate Conditions

I. Assemble full-length insert (tested for 16 parts, 9kb total)

[0136]

1.75 μ l	NF H ₂ O
4 μ l	80 nM total equimolar mixed parts (320 fmol)
1 μ l	10X 2.1 buffer NEB (10X CutSmart for BsaI)
1 μ l	1 mM rATP
2 μ l	BbsI NEB (10 U) (or 1 μ l BsaI, 10 U)
0.25 μ l	T4 DNA ligase HC Promega (5 U)
10 μ l total	

EP 3 149 168 B1

1. In thermal cycler: 2 min at 37°C, 5 min at 16°C for 20-24 hr (180 cycles in fast reaction module).
2. Inactivate at 50°C for 15 min, at 80°C for 10 min, and freeze at -20°C (ligation will continue if held at 0-4°C).
3. (Optional) Check assembly on a gel or capillary electrophoresis instrument.

II. Assemble insert into linear backbone (instead of destination vector with LacZ α to reduce background)

[0137] To prep linear backbone, inverse PCR is used to amplify the linear backbone from plasmid with flanking scars and Type IIs sites (should be the same Type IIs enzyme as in Part I above). Thoroughly digest with DpnI and gel purify the linear backbone before use to remove circular plasmid background.

8 μ l	16-part assembly reaction above
0.5 μ l	20 nM linear backbone with outer scars and Type IIs sites
0.25 μ l	1 mM rATP
1 μ l	NEB BbsI (or 0.25 μ l BsaI)
0.25 μ l	Promega T4 DNA ligase HC (5 U)

10 μ l Total

4. Set up backbone assembly. Include negative control with only the linear backbone for counting background CFU.
5. In thermal cycler: 4.5 hr at 37°C, then 15 min at 50°C, then 10min at 80°C, and then transform or freeze when done.
6. Transform 5 μ l of reaction into 50 μ l chemically competent cell aliquots. Transform negative control backbone only reaction. Add 100 μ l SOC and recover 1 hr. Plate entire recovery (optionally spin down cells to concentrate) on petri dish with appropriate antibiotic (optionally also include X-gal on plates as a control to check for the circular template from the backbone iPCR).
7. Ensure that there are more colonies for the part assembly reactions than for the backbone only negative control reaction. If not, the assembly reactions probably did not work and are only the empty vector.
8. Grow up colonies and digest the purified plasmid to check for insert.

[0138] Seventeen scars have been used, to date (listed below in order of use). A linear backbone with A, C flanking scars is used in certain embodiments.

List of Scars

[0139]

1	A	GGAG
2	D	AGGT
3	E	GCTT
4	F	CGCT
5	B	TACT
6	G	TGCC
7	H	GCAA
8	In	ACAG
9	Jn	ACGA
10	Kn	CAAC
11	L	CAGA
12	Mn	CCCA
13	Nn	CTTA
14	On	GATA
15	Pn	GTGA
16	Q	CCAT
17	c	AATG

Example 4

Assembly of Genetic Circuits

[0140] Multiplex pools were constructed containing 3,882,240 perfect match barcoded logic gate modules from a library of 10 repressors with 19 RBS-repressor variants of characterized NOT gate transfer functions. Our multiplex pooled library contains 35,746 unique gate modules, which corresponds to 98.5% of all possible NOT and 2-input NOR logic gates from our library of 10 repressors and 4 input sensors (Figure 50). From these pools, 10E26 different circuit designs could be constructed containing 1-10 gates with different wiring diagrams (permuted part assignments) and varied orientation (forward and reverse for each gate in any circuit). In total, 45% of sequencing reads were a perfect match to a designed gate module, which was improved from 20% in the about *nif* pathway multiplex retrievable libraries. It is believed this improvement is due to replacing the sequential type IIS scarless stitching and PCR with a one-pot type assembly and tagging reaction because in general these repressors are more toxic than the *nif* genes. Each repressor's NOT or NOR2 gate modules were constructed as a separate pool, and all 20 pools were sequenced in multiplex on one MiSeq v3 flow cell. On average there are 108 uniquely barcoded perfect match constructs for each designed gate module of the 35,746 unique gate modules constructed.

[0141] To design a set of logic circuit constructs, the Cello software was implemented with an additional clustering algorithm to select diverse part assignments to the abstract wiring diagram of gates. Initially, we choose two 3-input truth tables (0×40 and 0×89 in hexadecimal notation) and designed 10 different assignments for each. In total 77 different transcriptional units including the outputs were needed for these twenty designs. Six of the PBAD-containing transcriptional units were not present in the pooled library and were constructed individually from the parts library. The retrieved of 76 barcoded transcriptional unit modules from the pool were tested after optimizing the primer and template concentrations. Seventy-five of the retrieval PCRs (98.7%) amplified the product of the correct length (Figure 51), which exceeded our expectations given that 23 of these targets were singletons (identified from a single MiSeq read). Some of the retrieval PCRS showed multiple product lengths with additional undesired short bands. Therefore individual cloning and sequencing of a few of the multiple band retrieval PCRs were tested to confirm that only the full-length transcriptional unit assembles. The library of 20 circuit designs was assembled using a one-pot BbsI assembly, and verified full-length circuits for 15 of these designs on the initial attempt. These constructed logic circuits were tested by flow cytometry (YFP output) in all eight input states. Interestingly, most of these circuit designs result in the correct logic behavior in all 8 states. 24 constructed circuit plasmids (each gate -1.2 kb) were fully sequenced. It was found one error in 4 of the sequenced plasmids, yet two of these minor error-containing designs were functional.

Claims

1. A method of generating a set of sequence-verified nucleic acid elements for the combinatorial construction of a genetic design, the method comprising;

providing a plurality of nucleic acid parts, comprising one or more categories of nucleic acid parts, wherein nucleic acid parts of each category of parts share a defined structure and/or function, and wherein the plurality of nucleic acid parts comprise one or more of:

a set of nucleic acid sequences comprising different promoters;
a set of nucleic acid sequences comprising a set of different ribosomal binding sites;
a set of coding sequences; and/or
a set of nucleic acid sequences comprising a set of different transcription terminators;

assembling the nucleic acid parts to form a plurality of nucleic acid elements, wherein the nucleic acid elements comprise at least two sequences selected from the plurality of nucleic acid parts;

assembling onto the nucleic acid elements or parts a 3' and/or a 5' flanking nucleic acid sequence, wherein the 3' and/or 5' flanking sequences comprise one or more nucleic acid barcodes, such that each element of the set is uniquely tagged by the one or more nucleic acid barcodes, thereby facilitating selective retrieval of each individual nucleic acid element from the set; and

determining the sequence of the nucleic acid elements, thereby providing a set of sequence-verified nucleic acid elements.

2. The method of claim 1, wherein the barcodes comprise sequence specific priming sites to facilitate retrieval.

3. The method of claims 1 or 2, wherein the barcode comprises a DNA hash, comprising a unique nucleic acid identifier.
4. The method of claims 1-3, wherein the nucleic acid elements comprise a cistron or a subunit thereof.
5. The method of claims 1-4, wherein the nucleic acid elements comprise at least three sequences selected from the set of promoters, ribosomal binding sites, coding sequences, and transcription terminators, such as at least four sequences selected from the set of promoters, ribosomal binding sites, coding sequences and transcription terminators.
6. The method of claims 1-5, further comprising, retrieving one or more of the nucleic acid elements, such as wherein retrieval comprises nucleic acid amplification and/or isolation.
7. The method of claim 1-6, wherein:
 - a. the 3' and/or a 5' flanking nucleic acid sequence orders assembly of higher-order nucleic acid constructs;
 - b. the 3' and/or a 5' flanking nucleic acid sequence facilitates identification of the nucleic acid elements;
 - c. the nucleic acid parts further comprise one or more nucleic acid assembly sequences that order assembly;
 - d. the method is as defined in c) and the nucleic acid assembly sequence comprises one or more restriction sites, wherein the restriction site (s) are cleaved to produce sticky ends to facilitate ordered assembly of parts;
 - e. the 3' and/or 5' flanking sequences comprise one or more restriction sites, wherein the restriction site (s) are cleaved to produce sticky ends to facilitate ordered assembly of parts;
 - f. the method is as defined in d) or e) and the one or more restriction sites comprises at least one Type IIs restriction site;
 - g. the method is as defined in f) and cleavage of the Type IIs restriction site results in cohesive end overhangs at the 5' and 3' ends of the nucleic acid sequences such that the sequences are assembled into a continuous DNA fragment with the designed order and orientation of its components;
 - h. the nucleic acid parts further comprise one or more nucleic acid barcodes;
 - i. the method is as defined in claim 1 or in h) and the barcodes comprise a sequencing adaptor;
 - j. the method is as defined in claim 1 or in h) or i) and the barcode further comprises universal priming sites;
 - k. the method is as defined in claim 1 or 2 or in h) - j) and each of the barcodes further comprises a unique molecular identifier;
 - l. the method is as defined in claim 1 or 2 or in h) - k) and each of the barcodes comprises one or more indexes, one or more sequences that enable capture and/or amplification, and/or one or more sequences that enable library construction;
 - m. the method is as defined in claim 1-3 or in h) - l) and further comprises correlating the sequence of the barcode or hash to the sequence of the nucleic acid element or part;
 - n. two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same category of nucleic acid parts;
 - o. two or more of the assembled nucleic acid parts in the nucleic acid element are from the same category of nucleic acid parts;
 - p. the nucleic acid elements comprise a nucleic acid encoding a protein;
 - q. the method further comprises assembling a nucleic acid barcode to the 5' and/or 3' end of the nucleic acid element;
 - r. assembling comprises ligation of the nucleic acid sequences;
 - s. the assembly comprises Golden Gate Assembly and/or assembly using Type IIs restriction enzymes;
 - t. the assembly comprises Gibson Assembly, restriction ligation, ligation cycling reaction, ligation independent cloning, Gateway cloning, homologous recombination, and/or iterative capped assembly;
 - u. the method is as defined in claim 1 and determining the sequence comprises, nucleic acid sequencing, nucleic acid amplification, nucleic acid hybridization, or a combination thereof;
 - v. the nucleic acid parts comprise sequence variants, orthologs and/or homologs;
 - w. the method further comprises amplifying the nucleic acid elements to form a mixture of amplified set of nucleic acid elements;
 - x. the method is as defined in w) and amplification comprises polymerase chain reaction (PCR);
 - y. the method further comprises inserting the nucleic acid elements into a nucleic acid vectors, thereby generating a plurality of plasmids containing the nucleic acid elements;
 - z. the method is as defined in y) and the method further comprises pooling the plurality of plasmids;
 - aa. the method is as defined in claim 6 and the DNA molecule to retrieve is guided by a design algorithm;
 - bb. the method is as defined in claim 6 or in aa) and the design algorithm is one or more of the following: a

statistical algorithm, design of experiments (DOE), fractional factorial design, response surface methodology, and/or Taguchi method;

cc. the method is as defined in p) - bb) and the protein comprises an antibody or an antibody fragment;

dd. the method is as defined in cc) and the antibody or antibody fragment comprises a humanized antibody or antibody fragment;

ee. the method is as defined in cc) or dd) and the antibody or antibody fragment is of isotype IgG, IgA, IgM, IgE, or IgD, and/or the fragment selected comprises a Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP;

ff. the method is as defined in cc) - ee) and at least one of the nucleic acid parts encodes an antibody variable domain;

gg. at least one of the nucleic acid parts encodes an antibody constant domain;

hh. at least one of the nucleic acid parts encodes a protein from a biological pathway;

ii. the set of parts encodes multiple proteins from a biological pathway;

jj. the set of nucleic acid parts encodes all of the proteins from a biological pathway;

kk. the method is as defined in hh) - jj) and the biological pathway comprises a metabolic pathway;

ll. the method is as defined in kk) and the metabolic pathway comprises a synthetic pathway or a catabolic pathway;

mm. the method is as defined in hh) - ll) and the biological pathway comprises a genetic pathway; and/or

nn. the method is as defined in hh) - mm) and the biological pathway comprises a signal transduction pathway.

8. The method of any one of claims 1-7, further comprising generating a pool of higher-order nucleic acid constructs, the method further comprising:

retrieving two or more nucleic acid elements from the sequence-verified pool of nucleic acid elements; and assembling the two or more nucleic acid elements to form one or more higher-order nucleic acid constructs.

9. The method of claim 8, wherein:

a. the method further comprises assembling a nucleic acid barcode to the 5' and/or 3' end of the higher-order nucleic acid construct;

b. the assembly is sequential or concurrent;

c. assembling the two or more nucleic acid elements comprises ligation;

d. the assembly comprises Golden Gate Assembly and/or assembly using Type II restriction enzymes;

e. the assembly comprises Gibson Assembly, restriction ligation, ligation cycling reaction, ligation independent cloning, Gateway cloning, homologous recombination, and/or iterative capped assembly;

f. the method further comprises determining the sequence of the higher-order nucleic acid constructs thereby generating a pool of higher-order nucleic acid constructs;

g. the method is as defined in f) and determining the sequence comprises, nucleic acid sequencing, nucleic acid amplification, nucleic acid hybridization, or a combination thereof;

h. the method further comprises amplifying the higher-order nucleic acid constructs to form a mixture of amplified, higher-order nucleic acid constructs;

i. the method is as defined in h) and amplification comprises polymerase chain reaction (PCR);

j. the higher-order nucleic acid construct comprises two or more nucleic acid sequences encoding a protein;

k. the method is as defined in j) and the protein comprises an antibody or an antibody fragment;

l. the method is as defined in k) and the antibody or antibody fragment comprises a humanized antibody or antibody fragment;

m. the method is as defined in k) or l) and the antibody or antibody fragment is of isotype IgG, IgA, IgM, IgE, or IgD, and/or the fragment selected comprises a Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP;

n. the higher-order nucleic acid construct comprises two or more nucleic acid sequences encoding two or more proteins from a biological pathway;

o. the higher-order nucleic acid construct comprises two or more nucleic acid sequences encoding all of the proteins from a biological pathway;

p. the method is as defined in n) or o) and the biological pathway comprises a metabolic pathway;

q. the method is as defined in p) and the metabolic pathway comprises a synthetic pathway or a catabolic pathway;

r. the method is as defined in n) - q) and the biological pathway comprises a genetic pathway;

s. the method is as defined in n) - r) and the biological pathway comprises a signal transduction pathway;

t. the method further comprises inserting the higher-order nucleic acid constructs into a nucleic acid vectors, thereby generating a plurality of plasmids containing the higher-order nucleic acid constructs;

u. the method is as defined in t) and further comprises pooling the plurality of plasmids.

- v. the method is as defined in a) - u) and the barcodes comprise a sequencing adaptor;
- w. the method is as defined in a) - v) and the barcode further comprises universal priming sites;
- x. the method is as defined in a) - w) and the barcodes further comprise sequence specific priming sites to facilitate retrieval;
- y. the method is as defined in a) - x) and each of the barcodes further comprises a unique molecular identifier; and/or
- z. the method is as defined in a) - w) and each of the barcodes comprises one or more indexes, one or more sequences that enable capture and/or amplification, and/or one or more sequences that enable library construction.

10. The method of claims 1-9, further comprising:

assembling or incorporating a DNA hash, comprising a unique nucleic acid barcode to each of the plurality of nucleic acid parts, nucleic acid elements, cistrons, and/or higher-order nucleic acid constructs at the 5' and/or 3' end of the nucleic acid parts, nucleic acid elements, cistrons and/or higher-order nucleic acid constructs.

11. The method of claims 3-10, wherein:

- a. the method further comprises creating a hash table wherein the sequence of the individual unique nucleic acid barcode in the hash table represents the sequence of a nucleic acid part, nucleic acid elements, cistrons, and/or higher-order nucleic acid constructs;
- b. the method further comprises assigning the individual DNA hashes to the individual nucleic acid sequences in the plurality of nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs;
- c. the unique nucleic acid barcode comprises nucleic acid primer sites 5' and 3' of a unique molecular identifier;
- d. the unique molecular identifier is a unique DNA sequence 1-25 bp in length;
- e. the method further comprises amplifying the unique molecular identification sequence using primers that specifically hybridize to the nucleic acid primer sites;
- f. the method is as defined in e) and primers further comprise a nucleic acid sequence that specifically hybridizes to the sequence of one of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs;
- g. the method is as defined in e) or f) and further comprises assembling a concatemer of amplified unique molecular identification sequences.
- h. the method is as defined in g) and further comprises assembling the unique molecular identification sequences in an emulsion;
- i. the method is as defined in g) or h) and further comprises assembling the unique molecular identification sequences by overlap extension PCR, restriction ligation, ligation cycling reaction, type IIS assembly, homologous recombination, iterative capped assembly, and/or Gibson assembly; and/or
- j. the method is as defined in g) - i) and further comprises determining the sequence of the assembled concatemer, thereby determining the sequence of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs and/or the order of assembly of the nucleic acid parts, nucleic acid elements and/or higher-order nucleic acid constructs.

12. A set of sequence-verified nucleic acid elements for the combinatorial construction of genetic elements or amplification products thereof, comprising a plurality of nucleic acid elements assembled from a plurality of nucleic acid parts, comprising one or more categories of nucleic acid parts, wherein nucleic acid parts of each category of parts share a defined structure and/or function, and wherein the plurality of nucleic acid parts comprise one or more of: a set of nucleic acid sequences comprising different promoters; a set of nucleic acid sequences comprising a set of different ribosomal binding sites; a set of coding sequences; and/or a set of nucleic acid sequences comprising a set of different transcription terminators; wherein the nucleic acid elements comprise at least two sequences selected from the plurality of nucleic acid parts and a 3' and/or a 5' flanking nucleic acid sequence, wherein the 3' and/or 5' flanking sequences comprise one or more nucleic acid barcodes, such that each element of the set is uniquely tagged by the one or more nucleic acid barcodes, thereby facilitating selective retrieval of each individual nucleic acid element from the set.

13. The set of nucleic acid elements of claim 12, wherein:

- a. the 3' and/or a 5' flanking nucleic acid sequence orders assembly of higher-order nucleic acid constructs;
- b. the 3' and/or a 5' flanking nucleic acid sequence facilitates identification of the nucleic acid elements;
- c. the nucleic acid parts further comprise one or more nucleic acid assembly sequences that order assembly;
- d. the set is as defined in c) and the nucleic acid assembly sequence comprises one or more restriction sites

and, prior to assembly, they are cleaved to produce sticky ends to facilitate assembly;

e. the 3' and/or 5' flanking sequences comprise one or more restriction sites and, prior to assembly, they are cleaved to produce sticky ends to facilitate assembly;

f. the set is as defined in d) or e) and the one or more restriction sites comprises at least one Type IIs restriction site;

g. the set is as defined in f) and cleavage of the Type IIs restriction site results in cohesive end overhangs at the 5' and 3' ends of the nucleic acid sequences such that the sequences are assembled into a continuous DNA fragment with the designed order and orientation of its components;

h. the nucleic acid parts further comprise one or more nucleic acid barcodes;

i. the set is as defined in claim 12 or h) and the barcodes comprise a sequencing adaptor;

j. the set is as defined in claim 12 or h) and the barcode further comprises universal priming sites;

k. the set is as defined in h)-i) and the barcodes further comprise sequence specific priming sites to facilitate retrieval;

l. the set is as defined in h)-k) and each of the barcodes further comprises a unique molecular identifier;

m. the set is as defined in h)-l) and each of the barcodes comprises one or more indexes, one or more sequences that enable capture and/or amplification, and/or one or more sequences that enable library construction;

n. the set is as defined in h)-m) and the barcode comprises a DNA hash, comprising a unique nucleic acid identifier;

o. two or more of the assembled nucleic acid parts in the nucleic acid element are not from the same category of nucleic acid parts;

p. two or more of the assembled nucleic acid parts in the nucleic acid element are from the same category of nucleic acid parts;

q. the nucleic acid elements comprise a cistron or a subunit thereof;

r. the nucleic acid elements comprise a nucleic acid encoding a protein;

s. assembling comprises ligation of the nucleic acid sequences;

t. the sequence of the nucleic acid elements is determined;

u. the nucleic acid parts comprise sequence variants, orthologs and/or homologs;

v. the nucleic acid elements comprise at least three sequences selected from the set of a promoters, ribosomal binding sites, coding sequences, and transcription terminators;

w. the set is as defined in v) and the nucleic acid elements comprise at least four sequences selected from the set of a promoters, ribosomal binding sites, coding sequences and transcription terminators;

x. the set is as defined in r)-w) and the protein comprises an antibody or an antibody fragment;

y. the set is as defined in x) and the antibody or antibody fragment comprises a humanized antibody or antibody fragment;

z. the set is as defined in x)-y) and the antibody or antibody fragment is of isotype IgG, IgA, IgM, IgE, or IgD, and/or the fragment selected comprises a Fab, Fab', F(ab')₂, Fd, Fv, scFv, or SMIP;

aa. the set is as defined in x)-z) and at least one of the nucleic acid parts encodes an antibody variable domain;

bb. at least one of the nucleic acid parts encodes an antibody constant domain;

cc. at least one of the nucleic acid parts encodes a protein from a biological pathway;

dd. the set of parts encodes multiple proteins from a biological pathway;

ee. the set of nucleic acid parts encodes all of the proteins from a biological pathway;

ff. the set is as defined in cc)-ee) and the biological pathway comprises a metabolic pathway;

gg. the set is as defined in ff) and the metabolic pathway comprises a synthetic pathway or a catabolic pathway;

hh. the set is as defined in cc)-ff) and the biological pathway comprises a genetic pathway; and/or

ii. the set is as defined in cc)-hh) and the biological pathway comprises a signal transduction pathway;

14. A pool of higher-order nucleic acid constructs or amplification products thereof, comprising one or more nucleic acid elements assembled from the sequence-verified nucleic acid elements of claims 12-13.

15. A set of plasmids comprising the set of sequence-verified nucleic acid elements of claims 12-13 or comprising the pool of higher-order nucleic acid constructs of claim 14.

16. A kit comprising:

the pool of sequence-verified nucleic acid elements of claims 12-13 and/or the pool of higher-order nucleic acid constructs of claims 14; and a plurality of primers for retrieving one or more sequence-verified nucleic acid elements and/or higher-order nucleic acid constructs.

17. The set of nucleic acid elements of claim 12-13 or the pool of higher-order nucleic acid constructs of claim 14, wherein the sequence information for the set of sequence-verified nucleic acid elements and/or the pool of higher-

order nucleic acid constructs is comprised in an electronic file.

18. A method of optimizing a combination of nucleic acid elements for the production of an optimum desired phenotype, the method comprising

retrieving a plurality of sequence-verified nucleic acid elements of claims 12-13 and/or the pool of higher-order nucleic acid constructs of claims 14; assembling the retrieved sequences to construct a biological pathway assembly;
testing each of the assembled biological pathway assemblies for the desired phenotype;
determining which characteristic or characteristics of the assembled biological pathway contributes to the desired phenotype by comparing multiple assemblies;
based on the determination, plurality of sequence-verified nucleic acid elements of claims 12-13 and/or the pool of higher-order nucleic acid constructs of claims 14; assembling the retrieved sequences to construct a biological pathway assembly;
testing each of the assembled biological pathway assemblies for the desired phenotype; and repeating until an assembled biological pathway is identified that produces said desired phenotype.

Patentansprüche

1. Verfahren zum Erzeugen eines Satzes von sequenzverifizierten Nukleinsäureelementen für die kombinatorische Konstruktion eines genetischen Designs, wobei das Verfahren umfasst:

Bereitstellen mehrerer Nukleinsäureteile, umfassend eine oder mehrere Kategorien von Nukleinsäureteilen, wobei Nukleinsäureteile jeder Kategorie eine gemeinsame definierte Struktur und/oder Funktion aufweisen und wobei die mehreren Nukleinsäureteile eine oder mehrere der folgenden Komponenten umfassen:

einen Satz von Nukleinsäuresequenzen, die verschiedene Promotoren umfassen;
einen Satz von Nukleinsäuresequenzen, die einen Satz verschiedener ribosomaler Bindungsstellen umfassen;
einen Satz von Codierungssequenzen; und/oder
einen Satz von Nukleinsäuresequenzen, die einen Satz verschiedener Transkriptionsterminatoren umfassen;

Assemblieren der Nukleinsäureteile, um mehrere Nukleinsäureelemente auszubilden, wobei die Nukleinsäureelemente mindestens zwei Sequenzen umfassen, die aus den mehreren Nukleinsäureteilen ausgewählt sind; Assemblieren einer 3'- und/oder 5'-flankierenden Nukleinsäuresequenz an die Nukleinsäureelemente oder -teile, wobei die 3'- und/oder 5'-flankierenden Sequenzen einen oder mehrere Nukleinsäurebarcodes umfassen, sodass jedes Element des Satzes durch den einen oder die mehreren Nukleinsäurebarcodes eindeutig markiert wird, wodurch das selektive Abrufen jedes einzelnen Nukleinsäureelements aus dem Satz erleichtert wird; und Bestimmen der Sequenz der Nukleinsäureelemente, wodurch ein Satz von sequenzverifizierten Nukleinsäureelementen bereitgestellt wird.

2. Verfahren nach Anspruch 1, wobei die Barcodes sequenzspezifische Primingstellen umfassen, um das Abrufen zu erleichtern.
3. Verfahren nach Anspruch 1 oder 2, wobei der Barcode einen DNA-Hash umfasst, der eine eindeutige Nukleinsäureerkennung umfasst.
4. Verfahren nach den Ansprüchen 1-3, wobei die Nukleinsäureelemente ein Cistron oder eine Untereinheit davon umfassen.
5. Verfahren nach Anspruch 1-4, wobei die Nukleinsäureelemente mindestens drei Sequenzen umfassen, ausgewählt aus dem Satz von Promotoren, ribosomalen Bindungsstellen, Codierungssequenzen und Transkriptionsterminatoren, zum Beispiel mindestens vier Sequenzen ausgewählt aus dem Satz von Promotoren, ribosomalen Bindungsstelle, Codierungssequenzen und Transkriptionsterminatoren.
6. Verfahren nach Anspruch 1-5, ferner umfassend das Abrufen eines oder mehrerer der Nukleinsäureelemente, zum

Beispiel wenn das Abrufen eine Nukleinsäureamplifikation und/oder -isolation umfasst.

7. Verfahren nach Anspruch 1-6, wobei:

- a. die 3'- und/oder eine 5'-flankierende Nukleinsäuresequenz das Assemblieren von Nukleinsäurekonstrukten höherer Ordnung ordnet;
- b. die 3'- und/oder eine 5'-flankierende Nukleinsäuresequenz die Identifizierung der Nukleinsäureelemente ermöglicht;
- c. die Nukleinsäureteile ferner eine oder mehrere Nukleinsäureassemblierungssequenzen umfassen, die die Assemblierung ordnen;
- d. das Verfahren wie in c) definiert ist und die Nukleinsäureassemblierungssequenz eine oder mehrere Restriktionsstellen umfasst, wobei die Restriktionsstelle(n) gespalten werden, um klebrige Enden zu erzeugen, um die geordnete Assemblierung von Teilen zu erleichtern;
- e. die 3'- und/oder 5'-flankierenden Sequenzen eine oder mehrere Restriktionsstellen umfassen, wobei die Restriktionsstelle(n) gespalten werden, um klebrige Enden zu erzeugen, um die geordnete Assemblierung von Teilen zu erleichtern;
- f. das Verfahren wie in d) oder e) definiert ist und die eine oder die mehreren Restriktionsstellen mindestens eine Restriktionsstelle vom Typ IIs umfassen;
- g. das Verfahren wie in f) definiert ist und die Spaltung der Restriktionsstelle vom Typ IIs zu kohäsiven Endüberhängen an den 5'- und 3'-Enden der Nukleinsäuresequenzen führt, sodass die Sequenzen mit der entworfenen Reihenfolge und Ausrichtung ihrer Komponenten zu einem kontinuierlichen DNA-Fragment assembliert werden;
- h. die Nukleinsäureteile ferner einen oder mehrere Nukleinsäurebarcodes umfassen;
- i. das Verfahren wie in Anspruch 1 oder in h) definiert ist und die Barcodes einen Sequenzierungsadapter umfassen;
- j. das Verfahren wie in Anspruch 1 oder in h) oder i) definiert ist und der Barcode ferner universelle Primingstellen umfasst;
- k. das Verfahren wie in Anspruch 1 oder 2 oder in h)-j) definiert ist und jeder der Barcodes ferner eine eindeutige molekulare Kennung umfasst;
- l. das Verfahren wie in Anspruch 1 oder 2 oder in h)-k) definiert ist und jeder der Barcodes einen oder mehrere Indizes, eine oder mehrere Sequenzen, die die Erfassung und/oder Amplifikation ermöglichen, und/oder eine oder mehrere Sequenzen, die die Bibliotheksstruktur ermöglichen, umfasst;
- m. das Verfahren wie in Anspruch 1-3 oder in h)-l) definiert ist und ferner das Korrelieren der Sequenz des Barcodes oder des Hashs mit der Sequenz des Nukleinsäureelements oder -teils umfasst;
- n. zwei oder mehr der assemblierten Nukleinsäureteile im Nukleinsäureelement nicht zur gleichen Kategorie von Nukleinsäureteilen gehören;
- o. zwei oder mehr der assemblierten Nukleinsäureteile im Nukleinsäureelement aus derselben Kategorie von Nukleinsäureteilen stammen;
- p. die Nukleinsäureelemente eine Nukleinsäure umfassen, die für ein Protein kodiert;
- q. das Verfahren ferner das Assemblieren eines Nukleinsäurebarcodes an das 5'- und/oder das 3'-Ende des Nukleinsäureelements umfasst;
- r. Assemblieren das Ligieren der Nukleinsäuresequenzen umfasst;
- s. das Assemblieren Golden Gate-Assembly und/oder Assembly unter Verwendung von Restriktionsenzymen vom Typ IIs umfasst;
- t. das Assemblieren Gibson-Assembly, Restriktionsligation, Ligationszyklusreaktion, ligationsunabhängige Klonierung, Gateway-Klonierung, homologe Rekombination und/oder iteratives verkapptes Assemblieren umfasst;
- u. das Verfahren wie in Anspruch 1 definiert ist und das Bestimmen der Sequenz Nukleinsäuresequenzierung, Nukleinsäureamplifikation, Nukleinsäurehybridisierung oder eine Kombination davon umfasst;
- v. die Nukleinsäureteile Sequenzvarianten, Orthologe und/oder Homologe umfassen;
- w. das Verfahren ferner das Amplifizieren der Nukleinsäureelemente umfasst, um eine Mischung eines amplifizierten Satzes von Nukleinsäureelementen zu bilden;
- x. das Verfahren wie in w) definiert ist und die Amplifikation die Polymerasekettenreaktion (PCR) umfasst;
- y. das Verfahren ferner das Einfügen der Nukleinsäureelemente in einen Nukleinsäurevektor umfasst, wodurch mehrere Plasmide erzeugt werden, die die Nukleinsäureelemente enthalten;
- z. das Verfahren wie in y) definiert ist und das Verfahren ferner das Poolen der mehreren Plasmide umfasst;
- aa. das Verfahren wie in Anspruch 6 definiert ist und das abzurufende DNA-Molekül von einem Entwurfsalgorithmus geleitet wird;
- bb. das Verfahren wie in Anspruch 6 oder in aa) definiert ist und der Entwurfsalgorithmus einer oder mehrere

der folgenden ist: ein statistischer Algorithmus, Versuchsplanung (design of experiments - DOE), Teilfaktorplan, Antwortoberflächenmethode und/oder Taguchi-Verfahren;

cc. das Verfahren wie in p)-bb) definiert ist und das Protein einen Antikörper oder ein Antikörperfragment umfasst;

dd. das Verfahren wie in cc) definiert ist und der Antikörper oder das Antikörperfragment einen humanisierten Antikörper oder ein solches Antikörperfragment umfasst;

ee. das Verfahren wie in cc) oder dd) definiert ist und der Antikörper oder das Antikörperfragment den Isotyp IgG, IgA, IgM, IgE oder IgD aufweist, und/oder das ausgewählte Fragment ein Fab, Fab', F(ab')₂, Fd, Fv, scFv oder SMIP umfasst;

ff. das Verfahren wie in cc)-ee) definiert ist und mindestens einer der Nukleinsäureteile für eine variable Antikörperdomäne kodiert;

gg. mindestens einer der Nukleinsäureteile für eine konstante Antikörperdomäne kodiert;

hh. Mindestens einer der Nukleinsäureteile für ein Protein aus einem biologischen Signalweg kodiert;

ii. der Satz aus Teilen für mehrere Proteine aus einem biologischen Signalweg kodiert;

jj. der Satz aus Nukleinsäureteilen für alle Proteine aus einem biologischen Signalweg kodiert;

kk. das Verfahren wie in hh)-jj) definiert ist und der biologische Signalweg einen Stoffwechselweg umfasst;

ll. das Verfahren wie in kk) definiert ist und der Stoffwechselweg einen synthetischen Signalweg oder einen katabolischen Signalweg umfasst;

mm. das Verfahren wie in hh)-ll) definiert ist und der biologische Signalweg einen genetischen Signalweg umfasst; und/oder

nn. das Verfahren wie in hh)-mm) definiert ist und der biologische Signalweg einen Signaltransduktionsweg umfasst.

8. Verfahren nach einem der Ansprüche 1-7, ferner umfassend einen Pool von Nukleinsäurekonstrukten höherer Ordnung, wobei das Verfahren ferner umfasst:

Abrufen von zwei oder mehr Nukleinsäureelementen aus dem sequenzverifizierten Pool von Nukleinsäureelementen; und

Assemblieren der zwei oder mehr Nukleinsäureelemente, um ein oder mehrere Nukleinsäurekonstrukte höherer Ordnung auszubilden.

9. Verfahren nach Anspruch 8, wobei:

a. das Verfahren ferner das Assemblieren eines Nukleinsäurebarcodes an das 5'- und/oder das 3'-Ende des Nukleinsäurekonstrukts höherer Ordnung umfasst;

b. die Assemblierung sequenziell oder gleichzeitig stattfindet;

c. das Assemblieren der zwei oder mehr Nukleinsäureelemente Ligieren umfasst;

d. das Assemblieren Golden Gate-Assembly und/oder Assembly unter Verwendung von Restriktionsenzymen vom Typ IIs umfasst;

e. das Assemblieren Gibson-Assembly, Restriktionsligation, Ligationszyklusreaktion, ligationsunabhängige Klonierung, Gateway-Klonierung, homologe Rekombination und/oder iteratives verkapptes Assemblieren umfasst;

f. das Verfahren ferner das Bestimmen der Sequenz der Nukleinsäurekonstrukte höherer Ordnung umfasst, wodurch ein Pool von Nukleinsäurekonstrukten höherer Ordnung erzeugt wird;

g. das Verfahren wie in f) definiert ist und das Bestimmen der Sequenz Nukleinsäuresequenzierung, Nukleinsäureamplifikation, Nukleinsäurehybridisierung oder eine Kombination davon umfasst;

h. das Verfahren ferner das Amplifizieren der Nukleinsäurekonstrukte höherer Ordnung umfasst, um eine Mischung von amplifizierten Nukleinsäurekonstrukten höherer Ordnung auszubilden;

i. das Verfahren wie in h) definiert ist und die Amplifikation die Polymerasekettenreaktion (PCR) umfasst;

j. das Nukleinsäurekonstrukt höherer Ordnung zwei oder mehr Nukleinsäuresequenzen umfasst, die für ein Protein kodieren;

k. das Verfahren wie in j) definiert ist und das Protein einen Antikörper oder ein Antikörperfragment umfasst;

l. das Verfahren wie in k) definiert ist und der Antikörper oder das Antikörperfragment einen humanisierten Antikörper oder ein solches Antikörperfragment umfasst;

m. das Verfahren wie in k) oder 1) definiert ist und der Antikörper oder das Antikörperfragment den Isotyp IgG, IgA, IgM, IgE oder IgD aufweist, und/oder das ausgewählte Fragment ein Fab, Fab', F(ab')₂, Fd, Fv, scFv oder SMIP umfasst;

n. das Nukleinsäurekonstrukt höherer Ordnung zwei oder mehr Nukleinsäuresequenzen umfasst, die für zwei oder mehr Proteine aus einem biologischen Signalweg kodieren;

o. das Nukleinsäurekonstrukt höherer Ordnung zwei oder mehr Nukleinsäuresequenzen umfasst, die für alle

Proteine aus einem biologischen Signalweg kodieren;

p. das Verfahren wie in n) oder o) definiert ist und der biologische Signalweg einen Stoffwechselweg umfasst;

q. das Verfahren wie in p) definiert ist und der Stoffwechselweg einen synthetischen Signalweg oder einen katabolischen Signalweg umfasst;

r. das Verfahren wie in n)-q) definiert ist und der biologische Signalweg einen genetischen Signalweg umfasst;

s. das Verfahren wie in n)-r) definiert ist und der biologische Signalweg einen Signaltransduktionsweg umfasst;

t. das Verfahren ferner das Einfügen der Nukleinsäurekonstrukte höherer Ordnung in Nukleinsäurevektoren umfasst, wodurch mehrere Plasmide erzeugt werden, die die Nukleinsäurekonstrukte höherer Ordnung enthalten;

u. das Verfahren wie in t) definiert ist und ferner das Poolen der mehreren Plasmiden umfasst;

v. das Verfahren wie in a)-u) definiert ist und die Barcodes einen Sequenzierungsadapter umfassen;

w. das Verfahren wie in a)-v) definiert ist und der Barcode ferner universelle Primingstellen umfasst;

x. das Verfahren wie in a)-w) definiert ist und die Barcodes ferner sequenzspezifische Primingstellen umfassen, um das Abrufen zu ermöglichen;

y. das Verfahren wie in a)-x) definiert ist und jeder der Barcodes ferner eine eindeutige molekulare Kennung umfasst; und/oder

z. das Verfahren wie in a)-w) definiert ist und jeder der Barcodes einen oder mehrere Indizes, eine oder mehrere Sequenzen, die die Erfassung und/oder Amplifikation ermöglichen, und/oder eine oder mehrere Sequenzen, die die Bibliothekskonstruktion ermöglichen, umfasst.

10. Verfahren nach Anspruch 1-9, ferner umfassend:

Assemblieren oder Einarbeiten eines DNA-Hashs, umfassend einen eindeutigen Nukleinsäurebarcode für jede der mehreren Nukleinsäureteile, Nukleinsäureelemente, Cistrons und/oder Nukleinsäurekonstrukte höherer Ordnung am 5'- und/oder am 3'-Ende der Nukleinsäureteile, Nukleinsäureelemente, Cistrons und/oder Nukleinsäurekonstrukte höherer Ordnung.

11. Verfahren nach Anspruch 3-10, wobei:

a. das Verfahren ferner das Erstellen einer Hash-Tabelle umfasst, wobei die Sequenz des einzelnen eindeutigen Nukleinsäurebarcodes in der Hash-Tabelle die Sequenz eines Nukleinsäureteils, von Nukleinsäureelementen, Cistrons und/oder Nukleinsäurekonstrukten höherer Ordnung darstellt;

b. das Verfahren ferner das Zuordnen der einzelnen DNA-Hashes zu den einzelnen Nukleinsäuresequenzen der mehreren Nukleinsäureteile, Nukleinsäureelemente und/oder Nukleinsäurekonstrukte höherer Ordnung umfasst;

c. der eindeutige Nukleinsäurebarcode die Nukleinsäureprimerstellen 5' und 3' eines eindeutigen molekularen Identifikators umfasst;

d. der eindeutige molekulare Identifikator eine eindeutige DNA-Sequenz mit einer Länge von 1-25 bp ist;

e. das Verfahren ferner das Amplifizieren der eindeutigen molekularen Identifikationssequenz unter Verwendung von Primern umfasst, die spezifisch an die Nukleinsäureprimerstellen hybridisieren;

f. das Verfahren wie in e) definiert ist und Primer ferner eine Nukleinsäuresequenz umfassen, die spezifisch mit der Sequenz eines der Nukleinsäureteile, Nukleinsäureelemente und/oder Nukleinsäurekonstrukte höherer Ordnung hybridisiert;

g. das Verfahren wie in e) oder f) definiert ist und ferner das Assemblieren eines Concatemers amplifizierter eindeutiger molekularer Identifikationssequenzen umfasst;

h. das Verfahren wie in g) definiert ist und ferner das Assemblieren der eindeutigen molekularen Identifikationssequenzen in einer Emulsion umfasst;

i. das Verfahren wie in g) oder h) definiert ist und ferner das Assemblieren der eindeutigen molekularen Identifikationssequenzen durch Überlappungsverlängerungs-PCR, Restriktionsligation, Ligationszyklusreaktion, Typ IIS-Assemblierung, homologe Rekombination, iterative verkappte Assemblierung und/oder Gibson-Assemblierung umfasst; und/oder

j. das Verfahren wie in g)-i) definiert ist und ferner das Bestimmen der Sequenz des assemblierten Concatemers umfasst, wodurch die Sequenz der Nukleinsäureteile, Nukleinsäureelemente und/oder Nukleinsäurekonstrukte höherer Ordnung und/oder die Reihenfolge der Assemblierung der Nukleinsäureteile, Nukleinsäureelemente und/oder Nukleinsäurekonstrukte höherer Ordnung bestimmt wird.

12. Satz von sequenzverifizierten Nukleinsäureelementen für die kombinatorische Konstruktion genetischer Elemente oder von Amplifikationsprodukten davon, umfassend mehrere Nukleinsäureelemente, die assembliert sind aus meh-

5 reren Nukleinsäureteilen, umfassend eine oder mehrere Kategorien von Nukleinsäureteilen, wobei Nukleinsäureteile jeder Kategorie eine gemeinsame definierte Struktur und/oder Funktion aufweisen und wobei die mehreren Nukleinsäureteile eine oder mehrere der folgenden Komponenten umfassen: einen Satz von Nukleinsäuresequenzen, die verschiedene Promotoren umfassen; einen Satz von Nukleinsäuresequenzen, die einen Satz verschiedener ribosomaler Bindungsstellen umfassen; einen Satz von Codierungssequenzen; und/oder einen Satz von Nukleinsäuresequenzen, die einen Satz verschiedener Transkriptionsterminatoren umfassen; wobei die Nukleinsäureelemente mindestens zwei Sequenzen umfassen, ausgewählt aus den mehreren Nukleinsäureteilen und einer 3'- und/oder 5'-flankierenden Nukleinsäuresequenz, wobei die 3'- und/oder 5'-flankierenden Sequenzen einen oder mehrere Nukleinsäurebarcodes umfassen, sodass jedes Element des Satzes durch den einen oder die mehreren Nukleinsäurebarcodes eindeutig markiert wird, wodurch das selektive Abrufen jedes einzelnen Nukleinsäureelements aus dem Satz erleichtert wird.

13. Satz aus Nukleinsäureelementen nach Anspruch 12, wobei:

- a. die 3'- und/oder eine 5'-flankierende Nukleinsäuresequenz das Assemblieren von Nukleinsäurekonstrukten höherer Ordnung ordnet;
- b. die 3'- und/oder eine 5'-flankierende Nukleinsäuresequenz die Identifizierung der Nukleinsäureelemente ermöglicht;
- c. die Nukleinsäureteile ferner eine oder mehrere Nukleinsäureassemblierungssequenzen umfassen, die die Assemblierung ordnen;
- d. der Satz wie in c) definiert ist und die Nukleinsäureassemblierungssequenz eine oder mehrere Restriktionsstellen umfasst und diese vor der Assemblierung gespalten werden, um klebrige Enden zu erzeugen, um die Assemblierung zu erleichtern;
- e. die 3'- und/oder 5'-flankierenden Sequenzen eine oder mehrere Restriktionsstellen umfassen und diese vor der Assemblierung gespalten werden, um klebrige Enden zu erzeugen, um die Assemblierung zu erleichtern;
- f. der Satz wie in d) oder e) definiert ist und die eine oder die mehreren Restriktionsstellen mindestens eine Restriktionsstelle vom Typ IIs umfassen;
- g. der Satz wie in f) definiert ist und die Spaltung der Restriktionsstelle vom Typ IIs zu kohäsiven Endüberhängen an den 5'- und 3'-Enden der Nukleinsäuresequenzen führt, sodass die Sequenzen mit der entworfenen Reihenfolge und Ausrichtung ihrer Komponenten zu einem kontinuierlichen DNA-Fragment assembliert werden;
- h. die Nukleinsäureteile ferner einen oder mehrere Nukleinsäurebarcodes umfassen;
- i. der Satz wie in Anspruch 12 oder in h) definiert ist und die Barcodes einen Sequenzierungsadapter umfassen;
- j. der Satz wie in Anspruch 12 oder in h) definiert ist und der Barcode ferner universelle Primingstellen umfasst;
- k. der Satz wie in h)-i) definiert ist und die Barcodes ferner sequenzspezifische Primingstellen umfassen, um das Abrufen zu erleichtern;
- l. das Verfahren wie in h)-k) definiert ist und jeder der Barcodes ferner eine eindeutige molekulare Kennung umfasst;
- m. der Satz wie in h)-l) definiert ist und jeder der Barcodes einen oder mehrere Indizes, eine oder mehrere Sequenzen, die die Erfassung und/oder Amplifikation ermöglichen, und/oder eine oder mehrere Sequenzen, die die Bibliotheksstruktur ermöglichen, umfasst;
- n. der Satz wie in h)-m) definiert ist und der Barcode einen DNA-Hash umfasst, der einen eindeutigen Nukleinsäureidentifikator umfasst;
- o. zwei oder mehr der assemblierten Nukleinsäureteile im Nukleinsäureelement nicht zur gleichen Kategorie von Nukleinsäureteilen gehören;
- p. zwei oder mehr der assemblierten Nukleinsäureteile im Nukleinsäureelement aus derselben Kategorie von Nukleinsäureteilen stammen;
- q. die Nukleinsäureelemente ein Cistron oder eine Untereinheit davon umfassen;
- r. die Nukleinsäureelemente eine Nukleinsäure umfassen, die für ein Protein kodiert;
- s. Assemblieren das Ligieren der Nukleinsäuresequenzen umfasst;
- t. die Sequenz der Nukleinsäureelemente bestimmt wird;
- u. die Nukleinsäureteile Sequenzvarianten, Orthologe und/oder Homologe umfassen;
- v. die Nukleinsäureelemente mindestens drei Sequenzen umfassen, ausgewählt aus dem Satz von Promotoren, ribosomalen Bindungsstellen, Codierungssequenzen und Transkriptionsterminatoren;
- w. der Satz wie in v) definiert ist und die Nukleinsäureelemente mindestens vier Sequenzen umfassen, ausgewählt aus dem Satz von Promotoren, ribosomalen Bindungsstellen, Codierungssequenzen und Transkriptionsterminatoren;
- x. der Satz wie in r)-w) definiert ist und das Protein einen Antikörper oder ein Antikörperfragment umfasst;
- y. der Satz wie in x) definiert ist und der Antikörper oder das Antikörperfragment einen humanisierten Antikörper

oder ein solches Antikörperfragment umfasst;

z. der Satz wie in x)-y) definiert ist und der Antikörper oder das Antikörperfragment den Isotyp IgG, IgA, IgM, IgE oder IgD aufweist, und/oder das ausgewählte Fragment ein Fab, Fab', F(ab')₂, Fd, Fv, scFv oder SMIP umfasst;

aa. der Satz wie in x)-z) definiert ist und mindestens einer der Nukleinsäureteile für eine variable Antikörperdomäne kodiert;

bb. mindestens einer der Nukleinsäureteile für eine konstante Antikörperdomäne kodiert;

cc. Mindestens einer der Nukleinsäureteile für ein Protein aus einem biologischen Signalweg kodiert;

dd. der Satz aus Teilen für mehrere Proteine aus einem biologischen Signalweg kodiert;

ee. der Satz aus Nukleinsäureteilen für alle Proteine aus einem biologischen Signalweg kodiert;

ff. der Satz wie in cc)-ee) definiert ist und der biologische Signalweg einen Stoffwechselweg umfasst;

gg. der Satz wie in ff) definiert ist und der Stoffwechselweg einen synthetischen Signalweg oder einen katabolischen Signalweg umfasst;

hh. der Satz wie in cc)-ff) definiert ist und der biologische Signalweg einen genetischen Signalweg umfasst; und/oder

ii. der Satz wie in cc)-hh) definiert ist und der biologische Signalweg einen Signaltransduktionsweg umfasst.

14. Pool von Nukleinsäurekonstrukten höherer Ordnung oder von Amplifikationsprodukten davon, umfassend ein oder mehrere Nukleinsäureelemente, assembliert aus den sequenzverifizierten Nukleinsäure Elementen der Ansprüche 12-13.

15. Satz von Plasmiden, umfassend den Satz von sequenzverifizierten Nukleinsäureelementen nach Anspruch 12-13 oder umfassend den Pool von Nukleinsäurekonstrukten höherer Ordnung nach Anspruch 14.

16. Kit, umfassend:

den Pool von sequenzverifizierten Nukleinsäureelementen nach Anspruch 12-13 und/oder den Pool von Nukleinsäurekonstrukten höherer Ordnung nach Anspruch 14; und mehrere Primer zum Abrufen eines oder mehrerer sequenzverifizierter Nukleinsäureelemente und/oder Nukleinsäurekonstrukte höherer Ordnung.

17. Satz von Nukleinsäureelementen nach Anspruch 12-13 oder Pool von Nukleinsäurekonstrukten höherer Ordnung nach Anspruch 14, wobei die Sequenzinformationen für den Satz von sequenzverifizierten Nukleinsäureelementen und/oder für den Pool von Nukleinsäurekonstrukten höherer Ordnung in einer elektronischen Datei enthalten sind.

18. Verfahren zum Optimieren einer Kombination von Nukleinsäureelementen zum Erzeugen eines optimalen gewünschten Phänotyps, wobei das Verfahren umfasst:

Abrufen mehrerer sequenzverifizierter Nukleinsäureelemente nach Anspruch 12-13 und/oder des Pools von Nukleinsäurekonstrukten höherer Ordnung nach Anspruch 14; Assemblieren der abgerufenen Sequenzen, um eine biologische Signalweg-Assembly zu erzeugen;

Testen jeder der assemblierten biologischen Signalweg-Assemblies auf den gewünschten Phänotyp;

Bestimmen, welches Merkmal oder welche Merkmale des assemblierten biologischen Signalwegs zum gewünschten Phänotyp beitragen, indem mehrere Assemblies miteinander verglichen werden;

basierend auf dem Bestimmen, Abrufen mehrerer sequenzverifizierter Nukleinsäureelemente nach Anspruch 12-13 und/oder des Pools von Nukleinsäurekonstrukten höherer Ordnung nach Anspruch 14;

Assemblieren der abgerufenen Sequenzen, um eine biologische Signalweg-Assembly zu erzeugen;

Testen jeder der assemblierten biologischen Signalweg-Assemblies auf den gewünschten Phänotyp; und Wiederholen, bis ein assemblierter biologischer Signalweg identifiziert ist, der den gewünschten Phänotyp erzeugt.

Revendications

1. Procédé de génération d'un ensemble d'éléments d'acides nucléiques à séquence vérifiée destiné à la construction combinatoire d'une conception génétique, le procédé comprenant:

fournir une pluralité de parties d'acides nucléiques, comprenant une ou plusieurs catégorie(s) de parties d'acides nucléiques, dans lequel des parties d'acides nucléiques de chaque catégorie de parties partagent une structure et/ou une fonction définies, et dans lequel la pluralité de parties d'acides nucléiques comprend un ou plusieurs

parmi :

un ensemble de séquences d'acides nucléiques comprenant différents promoteurs ;
 un ensemble de séquences d'acides nucléiques comprenant un ensemble de sites de liaison ribosomiques différents ;
 un ensemble de séquences codantes ; et/ou
 un ensemble de séquences d'acides nucléiques comprenant un ensemble de terminateurs de la transcription différents ;

assembler les parties d'acides nucléiques pour former une pluralité d'éléments d'acides nucléiques, dans lequel les éléments d'acides nucléiques comprennent au moins deux séquences choisies dans la pluralité de parties d'acides nucléiques ;
 assembler, sur les éléments ou parties d'acides nucléiques, une séquence d'acides nucléiques flanquante en 3' et/ou en 5', dans lequel les séquences flanquantes en 3' et/ou en 5' comprennent un ou plusieurs code(s)-barres d'acides nucléiques, de sorte que chaque élément de l'ensemble soit étiqueté de manière unique par l'un ou les plusieurs code(s)-barres d'acides nucléiques, en facilitant de ce fait une récupération sélective de chaque élément d'acides nucléiques individuel à partir de l'ensemble ; et
 déterminer la séquence des éléments d'acides nucléiques, en fournissant de ce fait un ensemble d'éléments d'acides nucléiques à séquence vérifiée.

2. Procédé de la revendication 1, dans lequel les codes-barres comprennent des sites d'amorçage spécifiques d'une séquence pour faciliter une récupération.

3. Procédé des revendications 1 ou 2, dans lequel le code-barres comprend un hachage d'ADN, comprenant un identificateur d'acides nucléiques unique.

4. Procédé des revendications 1-3, dans lequel les éléments d'acides nucléiques comprennent un cistron ou une sous-unité de celui-ci.

5. Procédé des revendications 1-4, dans lequel les éléments d'acides nucléiques comprennent au moins trois séquences choisies parmi l'ensemble de promoteurs, de sites de liaison ribosomiques, de séquences codantes et de terminateurs de la transcription, tels qu'au moins quatre séquences choisies parmi l'ensemble de promoteurs, de sites de liaison ribosomiques, de séquences codantes et de terminateurs de la transcription.

6. Procédé des revendications 1-5, comprenant en outre, une récupération d'un ou de plusieurs des éléments d'acides nucléiques, tel que dans lequel la récupération comprend une amplification et/ou un isolement d'acides nucléiques.

7. Procédé des revendications 1-6, dans lequel :

a. la séquence d'acides nucléiques flanquante en 3' et/ou en 5' ordonnent un assemblage de constructions d'acides nucléiques d'ordres supérieurs ;

b. la séquence d'acides nucléiques flanquante en 3' et/ou une en 5' facilitent l'identification des éléments d'acides nucléiques ;

c. les parties d'acides nucléiques comprennent en outre une ou plusieurs séquence(s) d'acides nucléiques d'assemblage qui ordonne/ordonnent l'assemblage ;

d. le procédé est tel que défini dans le paragraphe c) et la séquence d'assemblage d'acides nucléiques comprend un ou plusieurs site(s) de restriction, dans lequel le(s) site(s) de restriction est/sont coupé(s) pour produire des extrémités collantes pour faciliter un assemblage ordonné des parties ;

e. les séquences d'acides nucléiques flanquantes en 3' et/ou en 5' comprennent un ou plusieurs site(s) de restriction, dans lesquelles le(s) site(s) de restriction est/sont coupé(s) pour produire des extrémités collantes pour faciliter un assemblage ordonné des parties ;

f. le procédé est tel que défini dans les paragraphes d) ou e) et le ou les plusieurs site(s) de restriction comprend/comprennent au moins un site de restriction de Type IIs ;

g. le procédé est tel que défini dans le paragraphe f) et une coupure des sites de restriction de Type IIs conduit à des surplombs d'extrémités cohésives au niveau des extrémités 5' et 3' des séquences d'acides nucléiques, de sorte que les séquences soient assemblées en un fragment d'ADN continu avec l'ordre et l'orientation conçus de ses composants ;

h. les parties d'acides nucléiques comprennent en outre un ou plusieurs code(s)-barres d'acides nucléiques ;

- i. le procédé est tel que défini dans la revendication 1 ou dans le paragraphe h) et les codes-barres comprennent un adaptateur de séquençage ;
- j. le procédé est tel que défini dans la revendication 1 ou dans les paragraphes h) ou i) et le code-barres comprend en outre des sites d'amorçage universels ;
- 5 k. le procédé est tel que défini dans les revendications 1 ou 2 ou dans les paragraphes h)-j) et chacun des codes-barres comprend en outre un identificateur moléculaire unique ;
- l. le procédé est tel que défini dans les revendications 1 ou 2 ou dans les paragraphes h)-k) et chacun des codes-barres comprend un ou plusieurs indice(s), une ou plusieurs séquence(s) qui permet/permettent de capturer et/ou d'amplifier, et/ou une ou plusieurs séquence(s) qui permette/permettent de construire une
- 10 banque ;
- m. le procédé est tel que défini dans les revendications 1-3 ou dans les paragraphes h)-1) et comprend en outre une corrélation de la séquence du code-barres ou du hachage avec la séquence de l'élément ou de la partie d'acides nucléiques ;
- n. deux des parties d'acides nucléiques assemblées ou plus dans l'élément d'acides nucléiques ne sont pas
- 15 de la même catégorie de parties d'acides nucléiques ;
- o. deux des parties d'acides nucléiques assemblées ou plus dans l'élément d'acides nucléiques sont de la même catégorie de parties d'acides nucléiques ;
- p. les éléments d'acides nucléiques comprennent un acide nucléique codant pour une protéine ;
- q. le procédé comprend en outre un assemblage d'un code-barres d'acides nucléiques avec l'extrémité 5' et/ou
- 20 3' de l'élément d'acides nucléiques ;
- r. un assemblage comprend une ligature des séquences d'acides nucléiques ;
- s. l'assemblage comprend un assemblage de type "Golden Gate" et/ou un assemblage utilisant des enzymes de restriction de Type IIs ;
- t. l'assemblage comprend un assemblage de "Gibson", une ligature de restriction, une réaction de cyclage de
- 25 ligature, un clonage indépendant de la ligature, un clonage "Gateway", une recombinaison homologue et/ou un assemblage à coiffe itérative ;
- u. le procédé est tel que défini dans la revendication 1 et une détermination de la séquence comprend, un séquençage d'acides nucléiques, une amplification d'acides nucléiques, une hybridation d'acides nucléiques ou une combinaison de ceux-ci ;
- 30 v. les parties d'acides nucléiques comprennent des variants, orthologues et/ou homologues de séquence ;
- w. le procédé comprend en outre une amplification des éléments d'acides nucléiques pour former un mélange d'un ensemble amplifié d'éléments d'acides nucléiques ;
- x. le procédé est tel que défini dans le paragraphe w) et une amplification comprend une réaction en chaîne par polymérase (PCR) ;
- 35 y. le procédé comprend en outre une insertion des éléments d'acides nucléiques dans des vecteurs d'acides nucléiques, en générant de ce fait une pluralité de plasmides contenant les éléments d'acides nucléiques ;
- z. le procédé est tel que défini dans le paragraphe y) et le procédé comprend en outre un regroupement de la pluralité de plasmides ;
- aa. le procédé est tel que défini dans la revendication 6 et la molécule d'ADN à récupérer est guidée par un
- 40 algorithme de conception ;
- bb. le procédé est tel que défini dans la revendication 6 ou dans le paragraphe aa) et l'algorithme de conception en est un ou plusieurs des suivants : un algorithme statistique, une conception d'expériences (DOE), une conception factorielle fractionnaire, une méthodologie de surface de réponse et/ou la méthode de Taguchi ;
- cc. le procédé est tel que défini dans les paragraphes p)-bb) et la protéine comprend un anticorps ou un fragment
- 45 d'anticorps ;
- dd. le procédé est tel que défini dans le paragraphe cc) et l'anticorps ou le fragment d'anticorps comprend un anticorps ou un fragment d'anticorps humanisé ;
- ee. le procédé est tel que défini dans les paragraphes cc) ou dd) et l'anticorps ou le fragment d'anticorps est de l'isotype IgG, IgA, IgM, IgE ou IgD et/ou le fragment choisi comprend un Fab, un Fab', un F(ab')₂, un Fd,
- 50 un Fv, un scFv ou un SMIP ;
- ff. le procédé est tel que défini dans les paragraphes cc)-ee) et au moins l'une des parties d'acides nucléiques code pour un domaine variable d'anticorps ;
- gg. au moins l'une des parties d'acides nucléiques code pour un domaine constant d'anticorps ;
- hh. au moins l'une des parties d'acides nucléiques code pour une protéine provenant d'une voie biologique ;
- 55 ii. l'ensemble de parties code pour de multiples protéines provenant d'une voie biologique ;
- jj. l'ensemble de parties d'acides nucléiques code pour toutes les protéines provenant d'une voie biologique ;
- kk. le procédé est tel que défini dans les paragraphes hh)-jj) et la voie biologique comprend une voie métabolique ;
11. le procédé est tel que défini dans le paragraphe kk) et la voie métabolique comprend une voie de synthèse

ou une voie catabolique ;

mm. le procédé est tel que défini dans les paragraphes hh)-ll) et la voie biologique comprend une voie génétique ;
et/ou

nn. le procédé est tel que défini dans les paragraphes hh)-mm) et la voie biologique comprend une voie de transduction du signal.

8. Procédé de l'une quelconque des revendications 1-7, comprenant en outre une génération d'un regroupement de constructions d'acides nucléiques d'ordres supérieurs, le procédé comprenant en outre:

recupérer deux éléments d'acides nucléiques ou plus à partir du regroupement à séquences vérifiées d'éléments d'acides nucléiques ; et
assembler les deux éléments d'acides nucléiques ou plus pour former une ou plusieurs construction(s) d'acides nucléiques d'ordre(s) supérieur(s).

9. Procédé de la revendication 8, dans lequel :

a. le procédé comprend en outre un assemblage d'un code-barres d'acides nucléiques au niveau de l'extrémité 5' et/ou 3' de la construction d'acides nucléiques d'ordre supérieur ;

b. l'assemblage est séquentiel ou concourant ;

c. un assemblage de deux éléments d'acides nucléiques ou plus comprend une ligature ;

d. l'assemblage comprend un assemblage de type "Golden Gate" et/ou un assemblage utilisant des enzymes de restriction de Type IIs ;

e. l'assemblage comprend un assemblage de "Gibson", une ligature de restriction, une réaction de cyclage de ligature, un clonage indépendant de la ligature, un clonage "Gateway", une recombinaison homologue et/ou un assemblage à coiffe itérative ;

f. le procédé comprend en outre une détermination de la séquence des constructions d'acides nucléiques d'ordres supérieurs en générant de ce fait un regroupement de constructions d'acides nucléiques d'ordres supérieurs ;

g. le procédé est tel que défini dans le paragraphe f) et une détermination de la séquence comprend, un séquençage d'acides nucléiques, une amplification d'acides nucléiques, une hybridation d'acides nucléiques ou une combinaison de ceux-ci ;

h. le procédé comprend en outre une amplification des constructions d'acides nucléiques d'ordres supérieurs pour former un mélange d'un ensemble de constructions d'acides nucléiques d'ordres supérieurs amplifiées ;

i. le procédé est tel que défini dans le paragraphe h) et une amplification comprend une réaction en chaîne par polymérase (PCR) ;

j. la construction d'acides nucléiques d'ordre supérieur comprend deux séquences d'acides nucléiques ou plus codant pour une protéine ;

k. le procédé est tel que défini dans le paragraphe j) et la protéine comprend un anticorps ou un fragment d'anticorps ;

l. le procédé est tel que défini dans le paragraphe k) et l'anticorps ou le fragment d'anticorps comprend un anticorps ou un fragment d'anticorps humanisé ;

m. le procédé est tel que défini dans les paragraphes k) ou l) et l'anticorps ou le fragment d'anticorps est de l'isotype IgG, IgA, IgM, IgE ou IgD et/ou le fragment choisi comprend un Fab, un Fab', un F(ab')₂, un Fd, un Fv, un scFv ou un SMIP ;

n. la construction d'acides nucléiques d'ordre supérieur comprend deux séquences d'acides nucléiques ou plus codant pour deux protéines ou plus provenant d'une voie biologique ;

o. la construction d'acides nucléiques d'ordre supérieur comprend deux séquences d'acides nucléiques ou plus codant pour toutes les protéines provenant d'une voie biologique ;

p. le procédé est tel que défini dans les paragraphes n) ou o) et la voie biologique comprend une voie métabolique ;

q. le procédé est tel que défini dans le paragraphe p) et la voie métabolique comprend une voie de synthèse ou une voie catabolique ;

r. le procédé est tel que défini dans les paragraphes n)-q) et la voie biologique comprend une voie génétique ;

s. le procédé est tel que défini dans les paragraphes n)-r) et la voie biologique comprend une voie de transduction du signal ;

t. le procédé comprend en outre une insertion des constructions d'acides nucléiques d'ordres supérieurs dans des vecteurs d'acides nucléiques, en générant de ce fait une pluralité de plasmides contenant les constructions d'acides nucléiques d'ordres supérieurs ;

u. le procédé est tel que défini dans le paragraphe t) et le procédé comprend en outre un regroupement de la pluralité de plasmides ;
 v. le procédé est tel que défini dans les paragraphes a)-u) et les codes-barres comprennent un adaptateur de séquençage ;
 5 w. le procédé est tel que défini dans les paragraphes a)-v) et le code-barres comprend en outre des sites d'amorçage universels ;
 x. le procédé est tel que défini dans les paragraphes a)-w) et les codes-barres comprennent en outre des sites d'amorçage spécifiques d'une séquences pour faciliter une récupération ;
 y. le procédé est tel que défini dans les paragraphes a)-x) et chacun des codes-barres comprend en outre un
 10 identificateur moléculaire unique ; et/ou
 z. le procédé est tel que défini dans les paragraphes a)-w) et chacun des codes-barres comprend un ou plusieurs indice (s), une ou plusieurs séquence(s) qui permet/permettent de capturer et/ou d'amplifier, et/ou une ou plusieurs séquence(s) qui permette/permettent de construire une banque.

15 **10.** Procédé des revendications 1-9, comprenant en outre les étapes consistant à :

assembler ou incorporer un hachage d'ADN, comprenant un code-barres d'acides nucléiques unique, avec chacune de la pluralité de parties d'acides nucléiques, d'éléments d'acides nucléiques, de cistrons et/ou de constructions d'acides nucléiques d'ordres supérieurs au niveau de l'extrémité 5' et/ou 3' des parties d'acides
 20 nucléiques, des éléments d'acides nucléiques, des cistrons et/ou des constructions d'acides nucléiques d'ordres supérieurs.

11. Procédé des revendications 3-10, dans lequel :

a. le procédé comprend en outre la création d'un tableau de hachage, dans lequel la séquence du code-barres d'acides nucléiques unique individuel dans le tableau de hachage représente la séquence d'une partie d'acides nucléiques, d'éléments d'acides nucléiques, de cistrons et/ou de constructions d'acides nucléiques d'ordres supérieurs ;
 b. le procédé comprend en outre l'attribution des hachages d'ADN individuels aux séquences d'acides nucléiques individuelles dans la pluralité de parties d'acides nucléiques, d'éléments d'acides nucléiques et/ou de construc-
 30 tions d'acides nucléiques d'ordres supérieurs ;
 c. le code-barres d'acides nucléiques unique comprend des sites d'amorces d'acides nucléiques en 5' et 3' d'un identificateur moléculaire unique ;
 d. l'identificateur moléculaire unique est une séquence d'ADN unique d'une longueur de 1-25 pb ;
 35 e. le procédé comprend en outre une amplification de la séquence d'identification moléculaire unique en utilisant des amorces qui s'hybrident spécifiquement avec les sites d'amorces d'acides nucléiques ;
 f. le procédé est tel que défini dans le paragraphe e) et les amorces comprennent en outre une séquence d'acides nucléiques qui s'hybride spécifiquement avec la séquence de l'une des parties d'acides nucléiques, des éléments d'acides nucléiques et/ou des constructions d'acides nucléiques d'ordres supérieurs ;
 40 g. le procédé est tel que défini dans les paragraphes e) ou f) et comprend en outre un assemblage d'un concatémère de séquences d'identification moléculaire uniques.
 h. le procédé est tel que défini dans le paragraphe g) et comprend en outre un assemblage des séquences d'identification moléculaire uniques dans une émulsion ;
 i. le procédé est tel que défini dans les paragraphes g) ou h) et comprend en outre un assemblage des séquences d'identification moléculaire unique par une PCR d'allongement de chevauchements, une ligature de restriction, une réaction de cyclage de ligature, un assemblage de type IIS, une recombinaison homologue, un assemblage à coiffe itérative et/ou un assemblage de "Gibson" ; et/ou
 45 j. le procédé est tel que défini dans les paragraphes g)-i) et comprend en outre une détermination de la séquence du concatémère assemblé, en déterminant de ce fait la séquence des parties d'acides nucléiques, d'éléments d'acides nucléiques et/ou de constructions d'acides nucléiques d'ordres supérieurs et/ou l'ordre de l'assemblage des parties d'acides nucléiques, des éléments d'acides nucléiques et/ou des constructions d'acides nucléiques d'ordres supérieurs.
 50

55 **12.** Ensemble d'éléments d'acides nucléiques à séquence vérifiée destiné à la construction combinatoire d'éléments génétiques ou de produits de l'amplification de ceux-ci, comprenant une pluralité d'éléments d'acides nucléiques assemblés à partir d'une pluralité de parties d'acides nucléiques, comprenant une ou plusieurs catégorie(s) de parties d'acides nucléiques, dans lequel des parties d'acides nucléiques de chaque catégorie de parties partagent une structure et/ou une fonction définies, et dans lequel la pluralité de parties d'acides nucléiques comprend un ou

plusieurs parmi : un ensemble de séquences d'acides nucléiques comprenant différents promoteurs ; un ensemble de séquences d'acides nucléiques comprenant un ensemble de sites de liaison ribosomiques différents ; un ensemble de séquences codantes ; et/ou un ensemble de séquences d'acides nucléiques comprenant un ensemble de terminateurs de la transcription différents ; dans lequel les éléments d'acides nucléiques comprennent au moins deux séquences choisies dans la pluralité de parties d'acides nucléiques et une séquence d'acides nucléiques flanquante en 3' et/ou une en 5', dans lequel les séquences flanquantes en 3' et/ou en 5' comprennent un ou plusieurs code(s)-barres d'acides nucléiques, de sorte que chaque élément de l'ensemble soit étiqueté de manière unique par l'un ou les plusieurs code(s)-barres d'acides nucléiques, en facilitant de ce fait une récupération sélective de chaque élément d'acides nucléiques individuel à partir de l'ensemble.

13. Ensemble d'éléments d'acides nucléiques de la revendication 12, dans lequel :

- a. la séquence d'acides nucléiques flanquante en 3' et/ou une en 5' ordonnent un assemblage de constructions d'acides nucléiques d'ordres supérieurs ;
- b. la séquence d'acides nucléiques flanquante en 3' et/ou une en 5' facilitent l'identification des éléments d'acides nucléiques ;
- c. les parties d'acides nucléiques comprennent en outre une ou plusieurs séquence(s) d'acides nucléiques d'assemblage qui ordonne/ordonnent un assemblage ;
- d. l'ensemble est tel que défini dans le paragraphe c) et la séquence d'assemblage d'acides nucléiques comprend un ou plusieurs site(s) de restriction et, avant l'assemblage, ils sont coupés pour produire des extrémités collantes pour faciliter l'assemblage ;
- e. les séquences flanquantes d'acides nucléiques en 3' et/ou en 5' comprennent un ou plusieurs site(s) de restriction et, avant l'assemblage, ils sont coupés pour produire des extrémités collantes pour faciliter l'assemblage ;
- f. l'ensemble est tel que défini dans les paragraphes d) ou e) et le ou les plusieurs site(s) de restriction comprend/comprennent au moins un site de restriction de Type IIs ;
- g. l'ensemble est tel que défini dans le paragraphe f) et une coupure du site de restriction de Type IIs conduit à des surplombs d'extrémités cohésives au niveau des extrémités 5' et 3' des séquences d'acides nucléiques, de sorte que les séquences soient assemblées en un fragment d'ADN continu avec l'ordre et l'orientation conçus de ses composants ;
- h. les parties d'acides nucléiques comprennent en outre un ou plusieurs code(s)-barres d'acides nucléiques ;
- i. l'ensemble est tel que défini dans la revendication 12 ou dans le paragraphe h) et les codes-barres comprennent un adaptateur de séquençage ;
- j. l'ensemble est tel que défini dans la revendication 12 ou dans le paragraphe h) et le code-barres comprend en outre des sites d'amorçage universels ;
- k. l'ensemble est tel que défini dans les paragraphes h)-i) et les codes-barres comprennent en outre des sites d'amorçage spécifiques d'une séquence pour faciliter une récupération ;
- l. l'ensemble est tel que défini dans les paragraphes h)-k) et chacun des codes-barres comprend en outre un identificateur moléculaire unique ;
- m. l'ensemble est tel que défini dans les paragraphes h)-l) et chacun des codes-barres comprend un ou plusieurs indice(s), une ou plusieurs séquence(s) qui permet/permettent de capturer et/ou d'amplifier, et/ou une ou plusieurs séquence(s) qui permette/permettent de construire une banque ;
- n. l'ensemble est tel que défini dans les paragraphes h)-m) et le code-barres comprend un hachage d'ADN, comprenant un identificateur d'acides nucléiques unique ;
- o. deux des parties d'acides nucléiques assemblées ou plus dans l'élément d'acides nucléiques ne sont pas de la même catégorie de parties d'acides nucléiques ;
- p. deux des parties d'acides nucléiques assemblées ou plus dans l'élément d'acides nucléiques sont de la même catégorie de parties d'acides nucléiques ;
- q. les éléments d'acides nucléiques comprennent un cistron ou une sous-unité de celui-ci ;
- r. les éléments d'acides nucléiques comprennent un acide nucléique codant pour une protéine ;
- s. un assemblage comprend une ligature des séquences d'acides nucléiques ;
- t. la séquence des éléments d'acides nucléiques est déterminée ;
- u. les parties d'acides nucléiques comprennent des variants, orthologues et/ou homologues de séquence ;
- v. les éléments d'acides nucléiques comprennent au moins trois séquences choisies parmi l'ensemble de promoteurs, de sites de liaison ribosomiques, de séquences codantes et de terminateurs de la transcription ;
- w. l'ensemble est tel que défini dans le paragraphe v) et les éléments d'acides nucléiques comprennent au moins quatre séquences choisies parmi l'ensemble de promoteurs, de sites de liaison ribosomiques, de séquences codantes et de terminateurs de la transcription ;

x. l'ensemble est tel que défini dans les paragraphes r)-w) et la protéine comprend un anticorps ou un fragment d'anticorps ;

y. l'ensemble est tel que défini dans le paragraphe x) et l'anticorps ou le fragment d'anticorps comprend un anticorps ou un fragment d'anticorps humanisé ;

z. l'ensemble est tel que défini dans les paragraphes x)-y) et l'anticorps ou le fragment d'anticorps est de l'isotype IgG, IgA, IgM, IgE ou IgD et/ou le fragment choisi comprend un Fab, un Fab', un F(ab')₂, un Fd, un Fv, un scFv ou un SMIP ;

aa. l'ensemble est tel que défini dans les paragraphes x)-z) et au moins l'une des parties d'acides nucléiques code pour un domaine variable d'anticorps ;

bb. au moins l'une des parties d'acides nucléiques code pour un domaine constant d'anticorps ;

cc. au moins l'une des parties d'acides nucléiques code pour une protéine provenant d'une voie biologique ;

dd. l'ensemble de parties code pour de multiples protéines provenant d'une voie biologique ;

ee. l'ensemble de parties code pour toutes les protéines provenant d'une voie biologique ;

ff. l'ensemble est tel que défini dans les paragraphes cc)-ee) et la voie biologique comprend une voie métabolique ;

gg. l'ensemble est tel que défini dans le paragraphe ff) et la voie métabolique comprend une voie de synthèse ou une voie catabolique ;

hh. l'ensemble est tel que défini dans les paragraphes cc)-ff) et la voie biologique comprend une voie génétique ; et/ou

ii. l'ensemble est tel que défini dans les paragraphes cc)-hh) et la voie biologique comprend une voie de transduction du signal ;

14. Regroupement de constructions d'acides nucléiques d'ordres supérieurs ou de produits de l'amplification de celles-ci, comprenant un ou plusieurs élément(s) d'acides nucléiques assemblé(s) à partir des éléments d'acides nucléiques à séquence vérifiée des revendications 12-13.

15. Ensemble de plasmides comprenant l'ensemble des éléments d'acides nucléiques à séquence vérifiée des revendications 12-13 ou comprenant le regroupement de constructions d'acides nucléiques d'ordres supérieurs de la revendication 14.

16. Trousse comprenant :

le regroupement des éléments d'acides nucléiques à séquence vérifiée des revendications 12-13 et/ou le regroupement de constructions d'acides nucléiques d'ordres supérieurs de la revendication 14 ; et une pluralité d'amorces destinées à récupérer un ou plusieurs élément(s) d'acides nucléiques à séquence vérifiée et/ou construction(s) d'acides nucléiques d'ordre(s) supérieur (s) .

17. Ensemble d'éléments d'acides nucléiques des revendications 12-13 ou regroupement de constructions d'acides nucléiques d'ordres supérieurs de la revendication 14, dans lequel l'information de séquence pour l'ensemble d'éléments d'acides nucléiques à séquence vérifiée et/ou le regroupement de constructions d'acides nucléiques d'ordres supérieurs est comprise dans un fichier électronique.

18. Procédé d'optimisation d'une combinaison d'éléments d'acides nucléiques destinée à la production d'un phénotype optimum souhaité, le procédé comprenant:

récupérer une pluralité d'éléments d'acides nucléiques à séquence vérifiée des revendications 12-13 et/ou le regroupement de constructions d'acides nucléiques d'ordres supérieurs de la revendication 14 ; assembler les séquences récupérées pour construire un assemblage de voies biologiques ;

tester chacun des assemblages de voies biologiques assemblées en vue du phénotype souhaité ;

déterminer quelle caractéristique ou quelles caractéristiques de la voie biologique assemblée contribue au phénotype souhaité par une comparaison de multiples assemblages ;

sur la base de la détermination, d'une pluralité d'éléments d'acides nucléiques à séquence vérifiée des revendications 12-13 et/ou du regroupement de constructions d'acides nucléiques d'ordres supérieurs de la revendication 14 ; assembler les séquences récupérées pour construire un assemblage de voies biologiques ;

tester chacun des assemblages de voies biologiques assemblées en vue du phénotype souhaité ; et répéter jusqu'à qu'une voie biologique assemblée soit identifiée, laquelle produit ledit phénotype souhaité.

- ▶ **Combinatorial searches**
(16+ Dimensions)
- ▶ **Larger libraries of permuted designs**
(500-1,000's per design cycle)
- ▶ **Specified permutations per design**
(100+ parts/design)
- ▶ **High-fidelity, sequence-verified designs**
(Mb assembled and 1,000's parts)

FIGURE 1

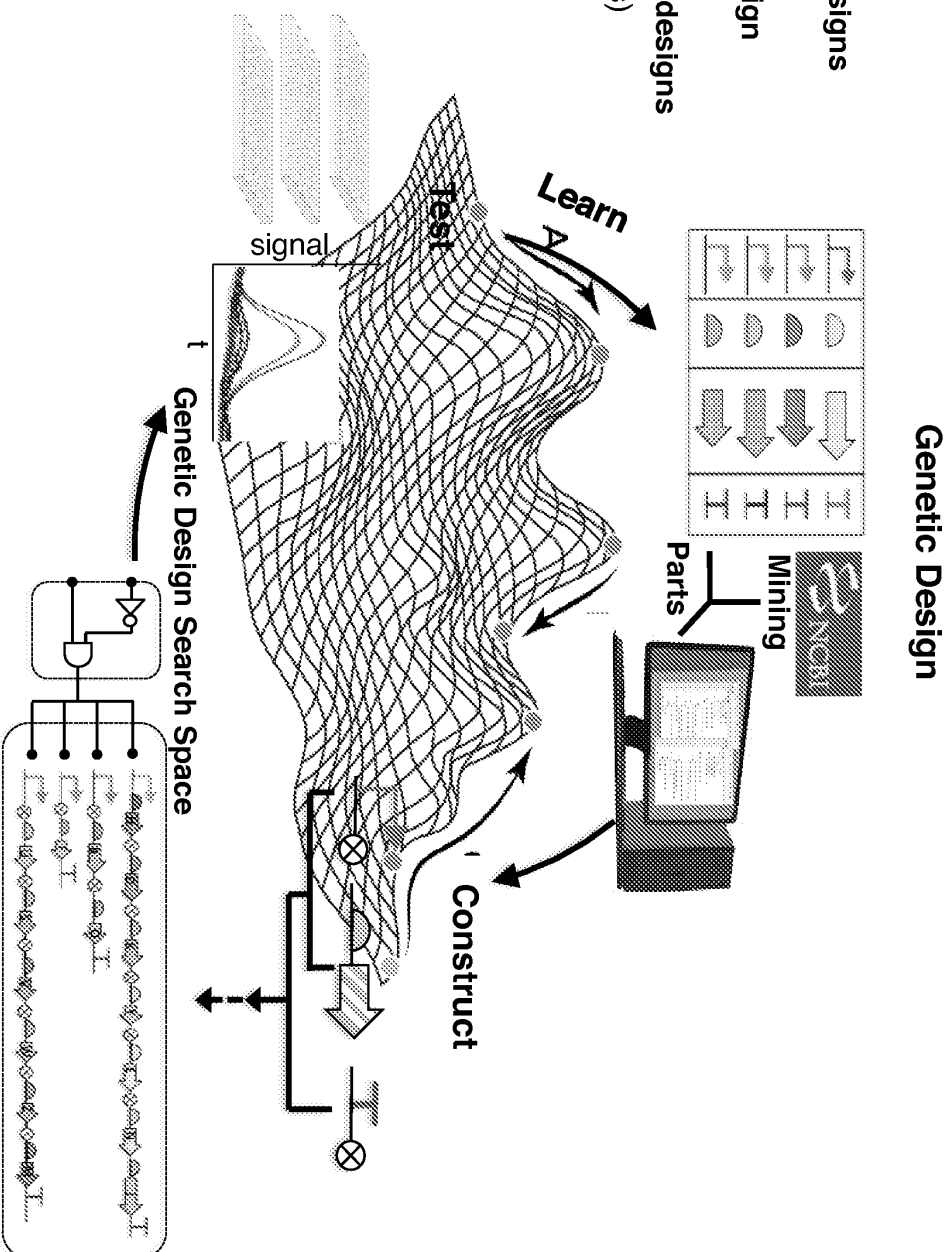
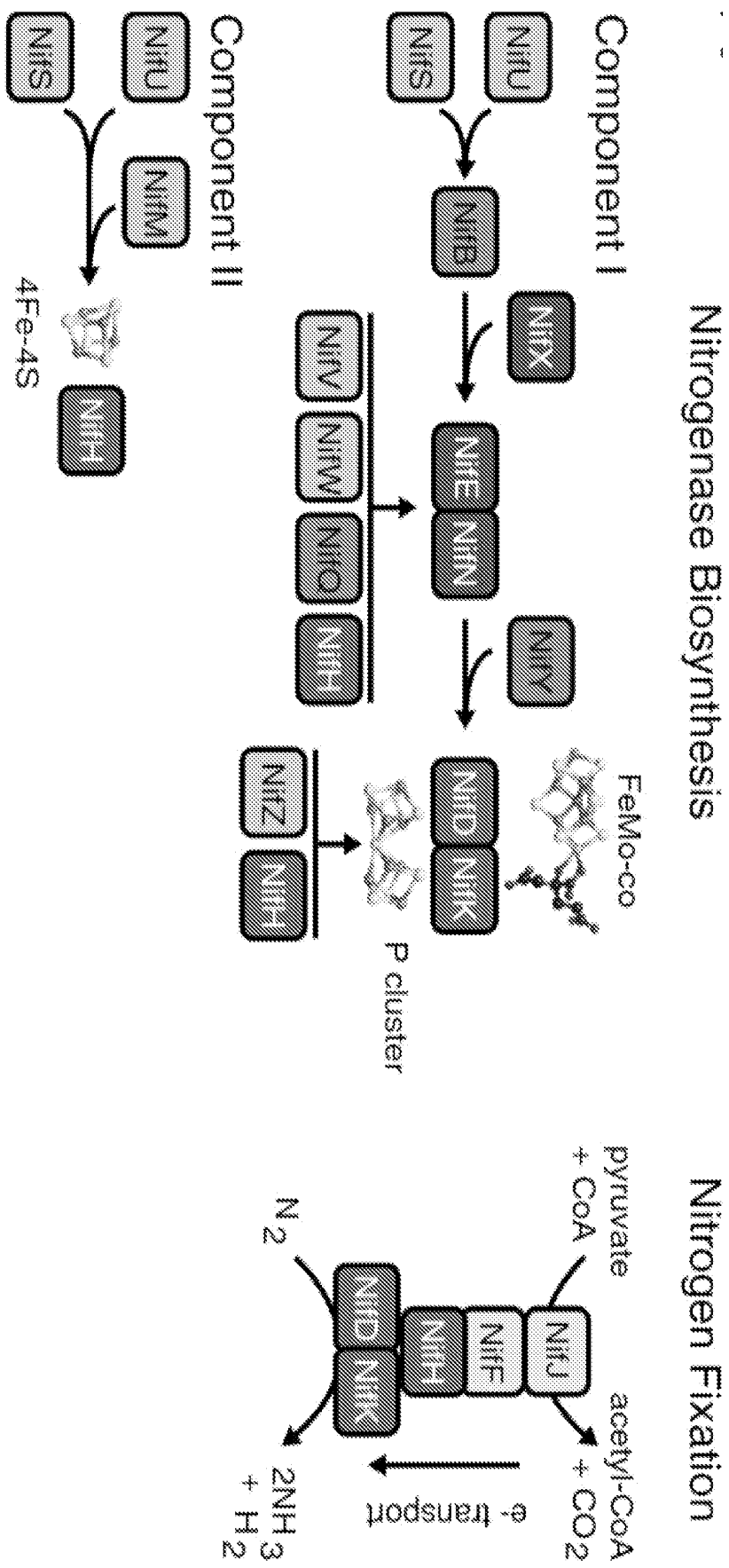


FIGURE 2



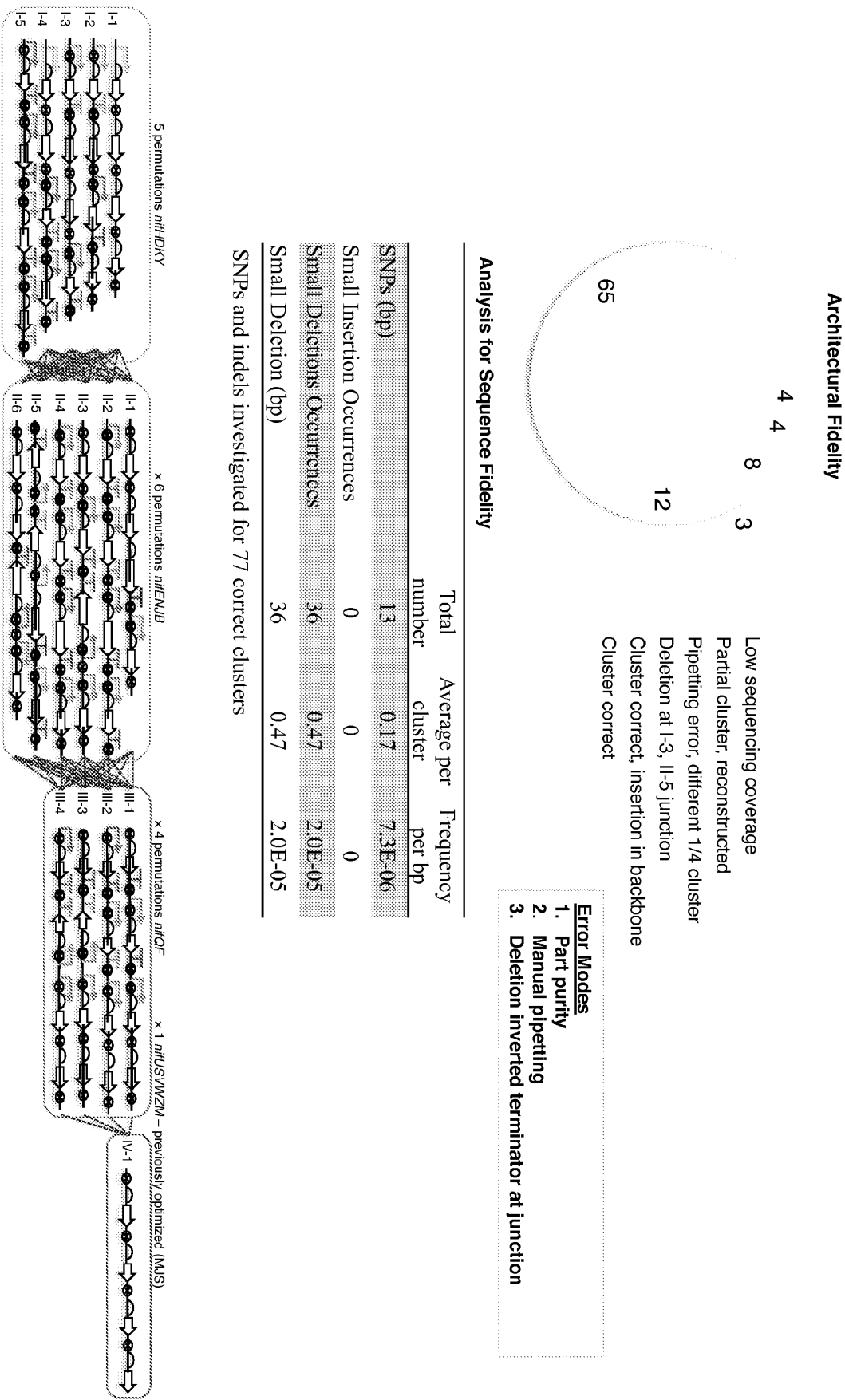
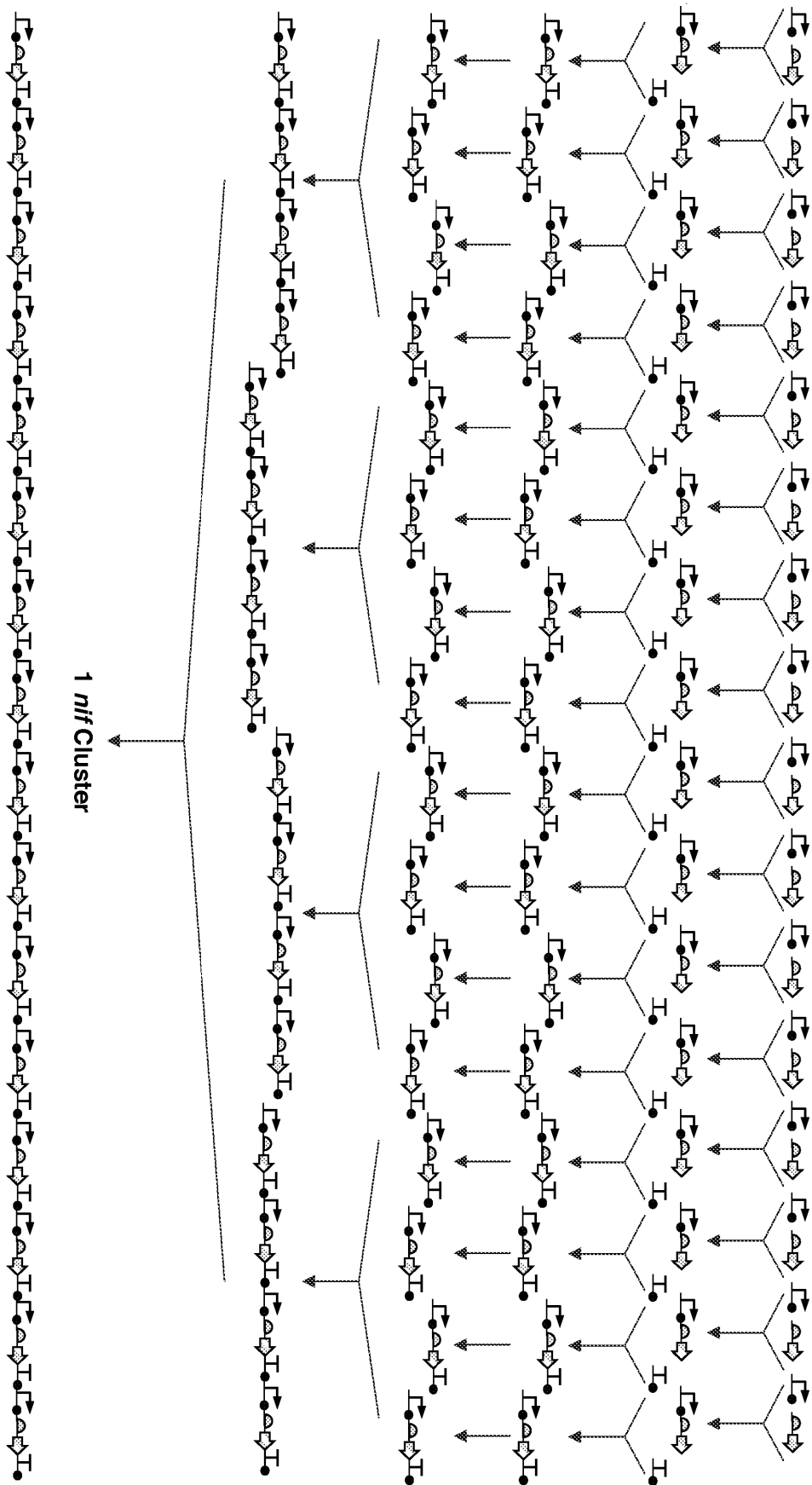


FIGURE 4



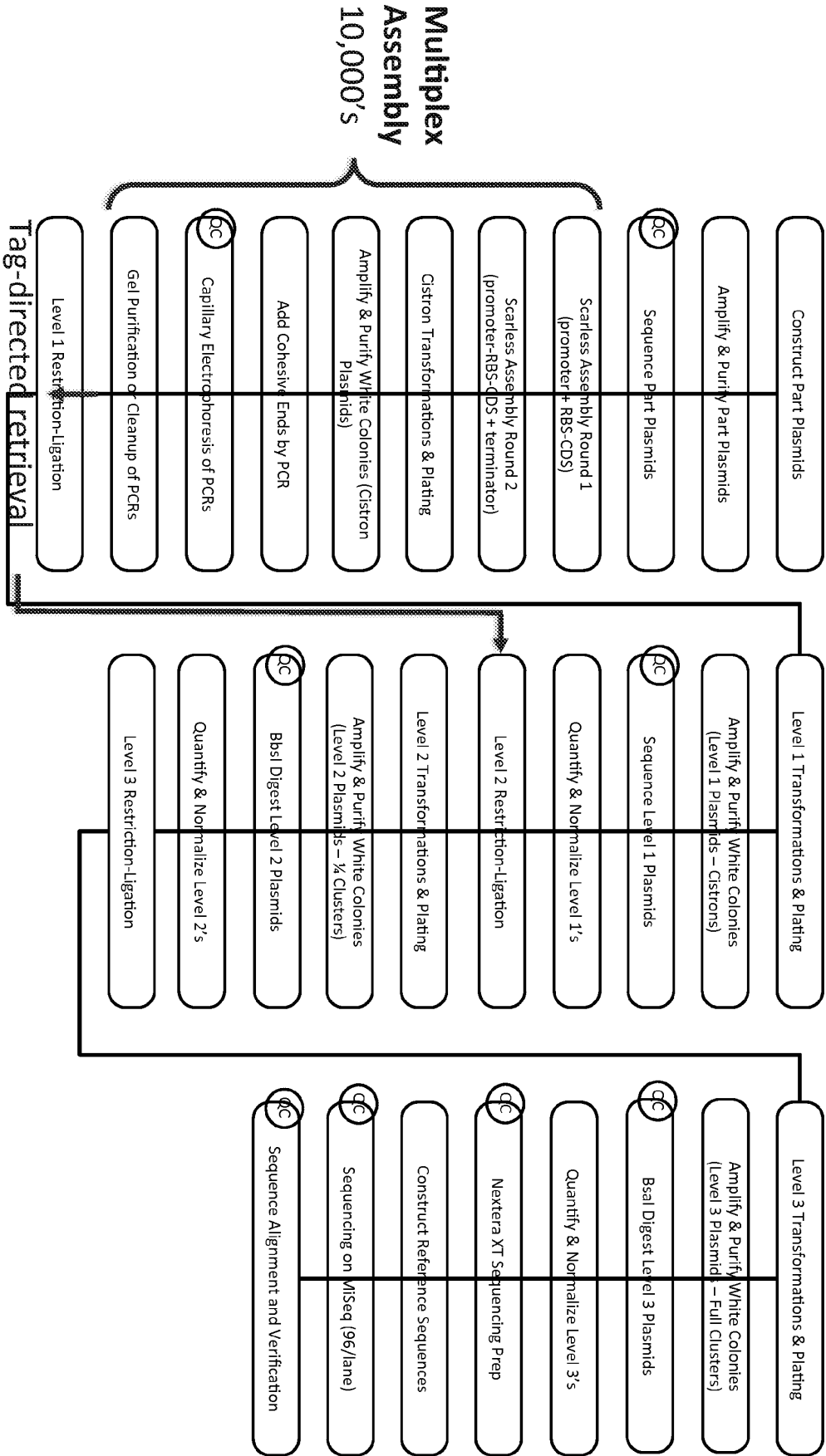


FIGURE 5

FIGURE 6

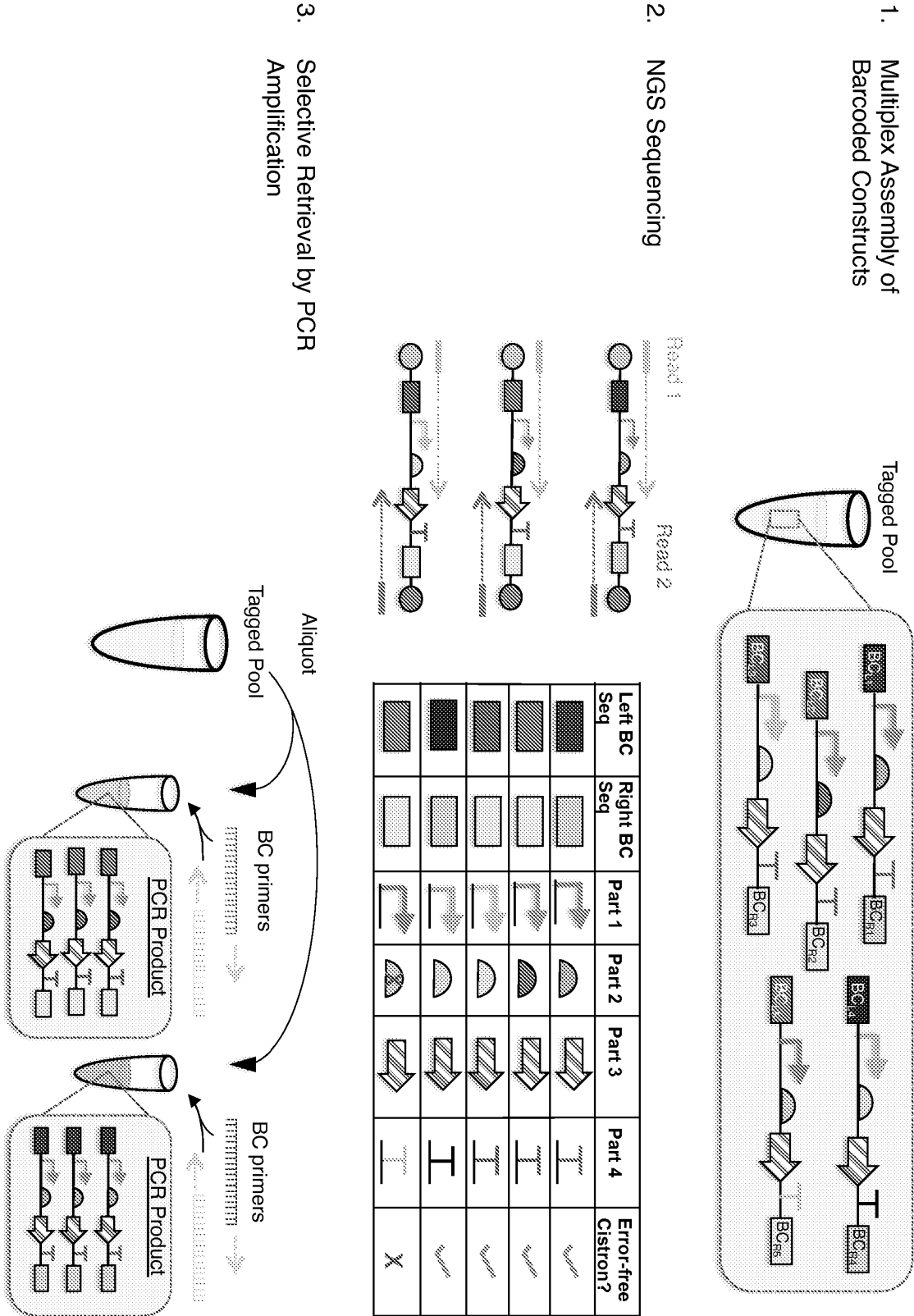


FIGURE 7

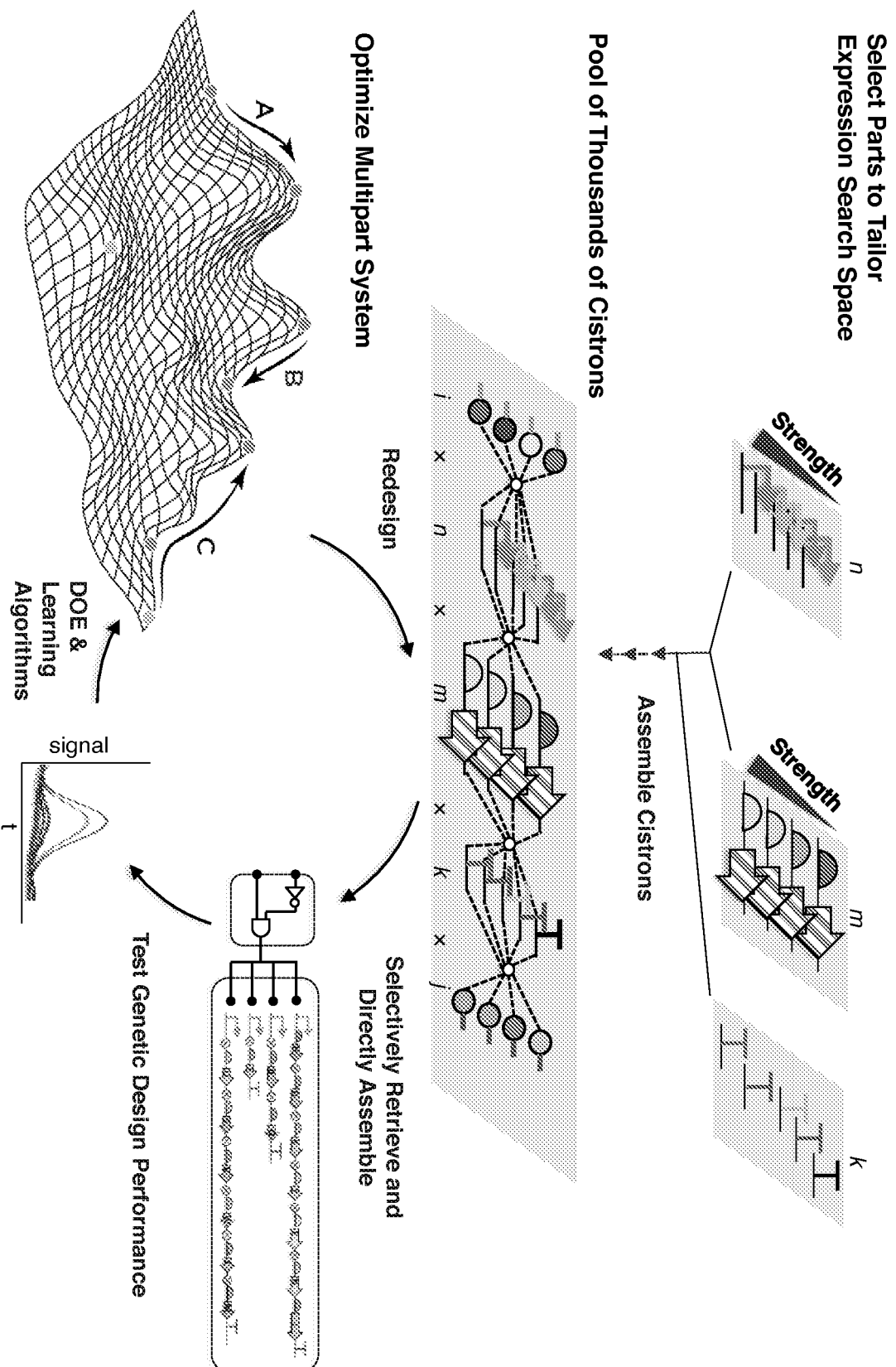


FIGURE 8

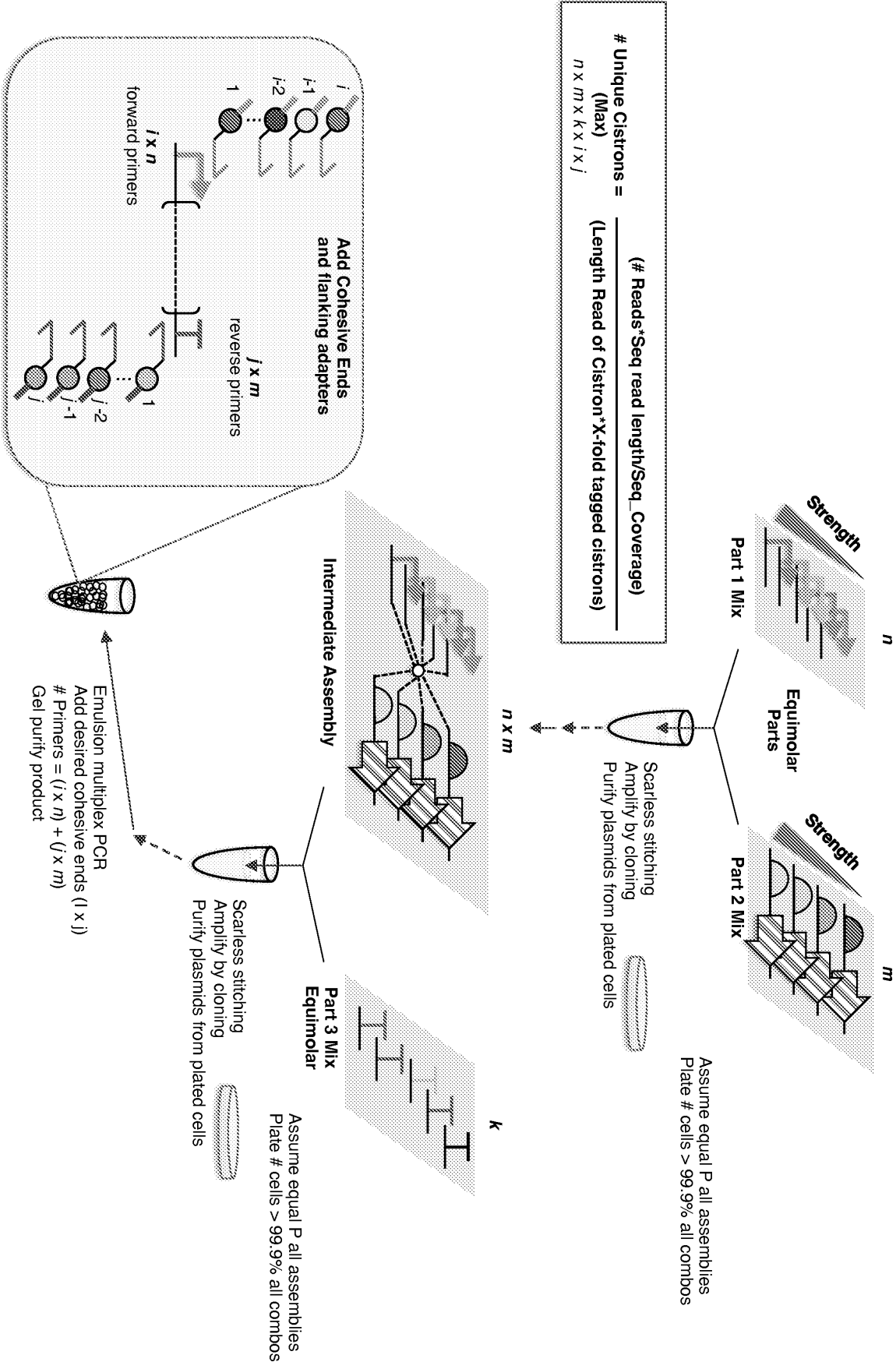


FIGURE 9

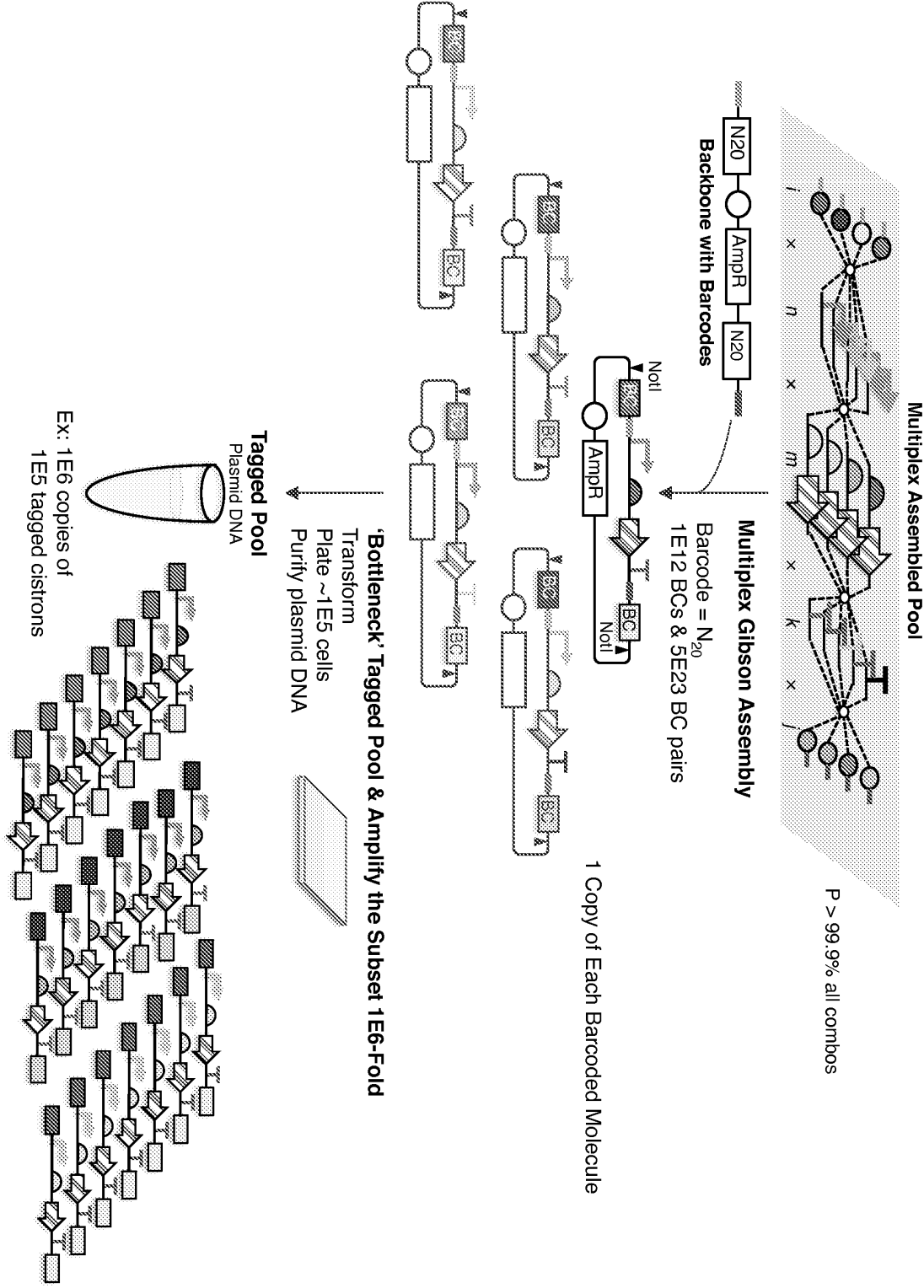


FIGURE 10

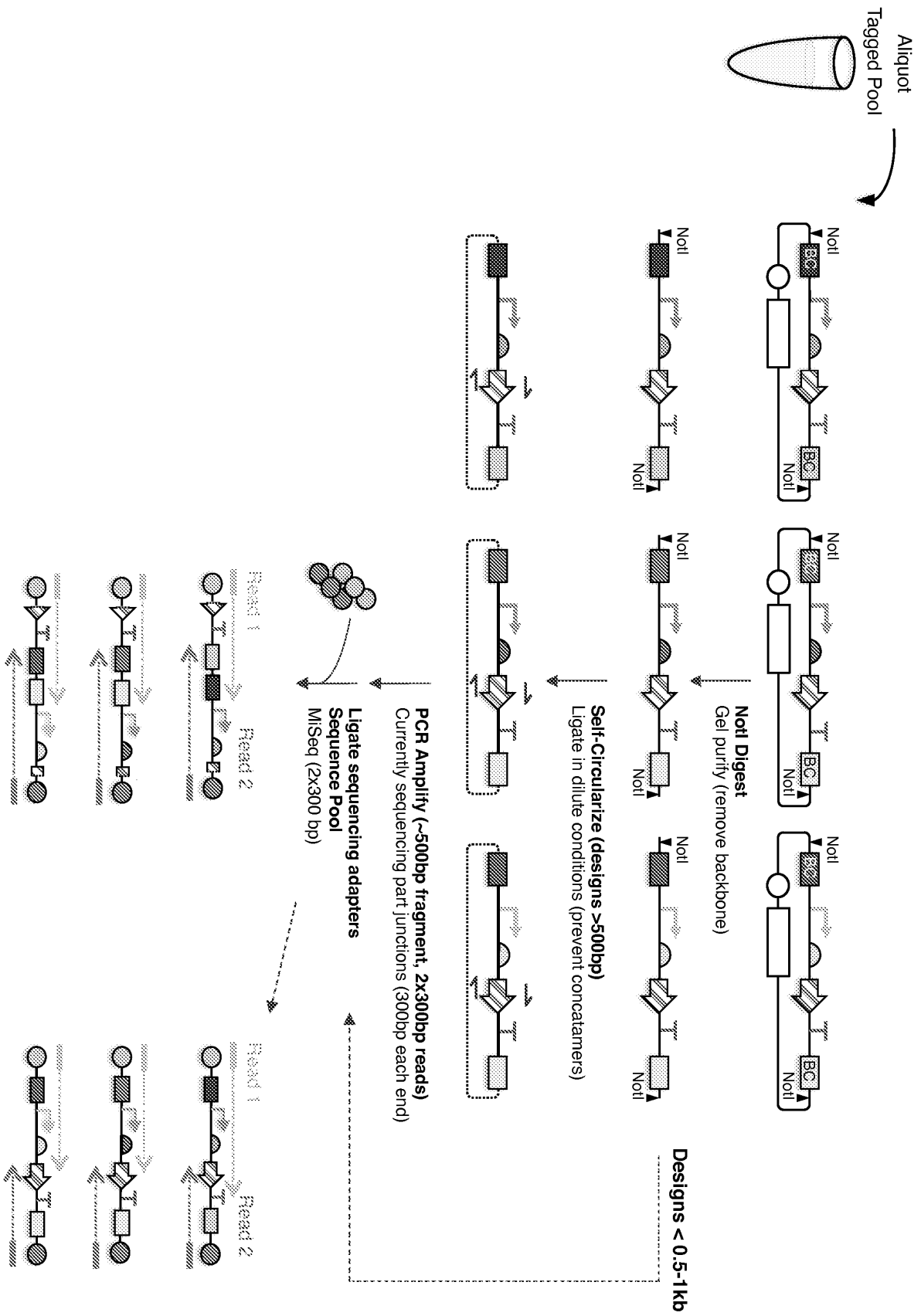


FIGURE 11

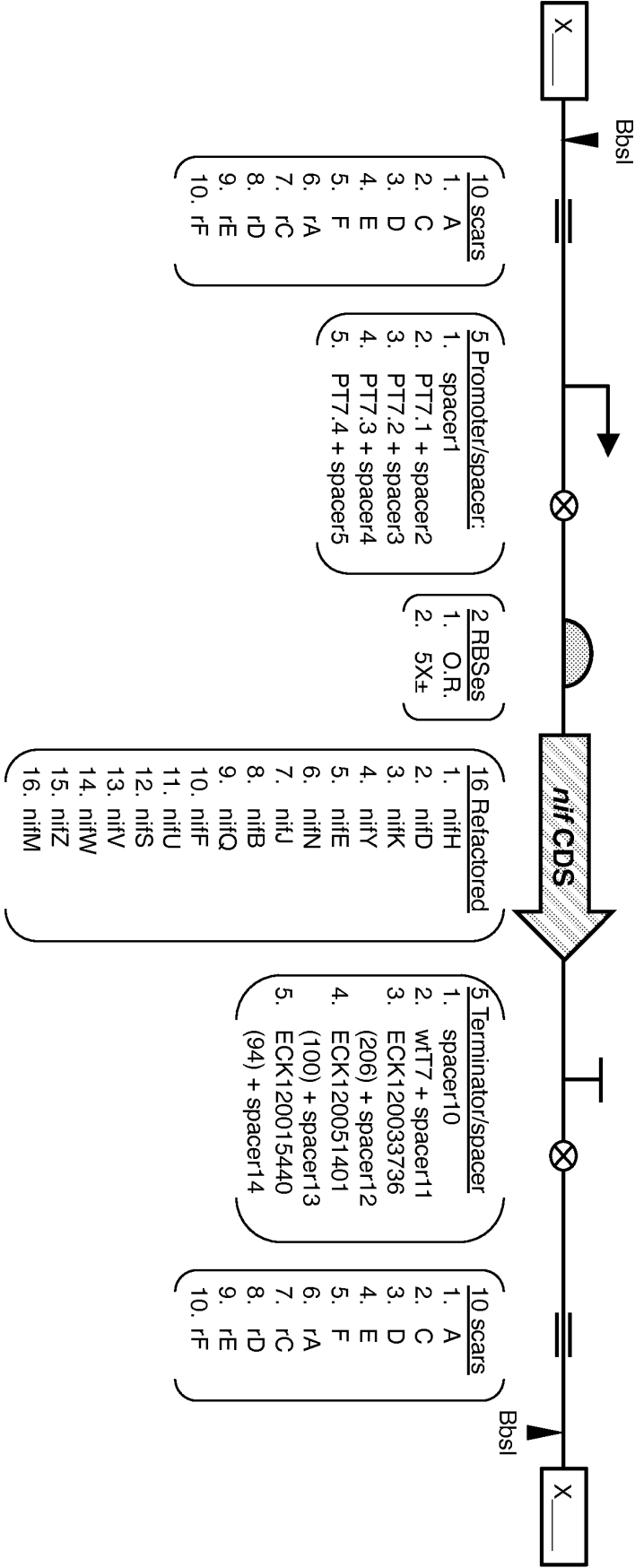


FIGURE 12

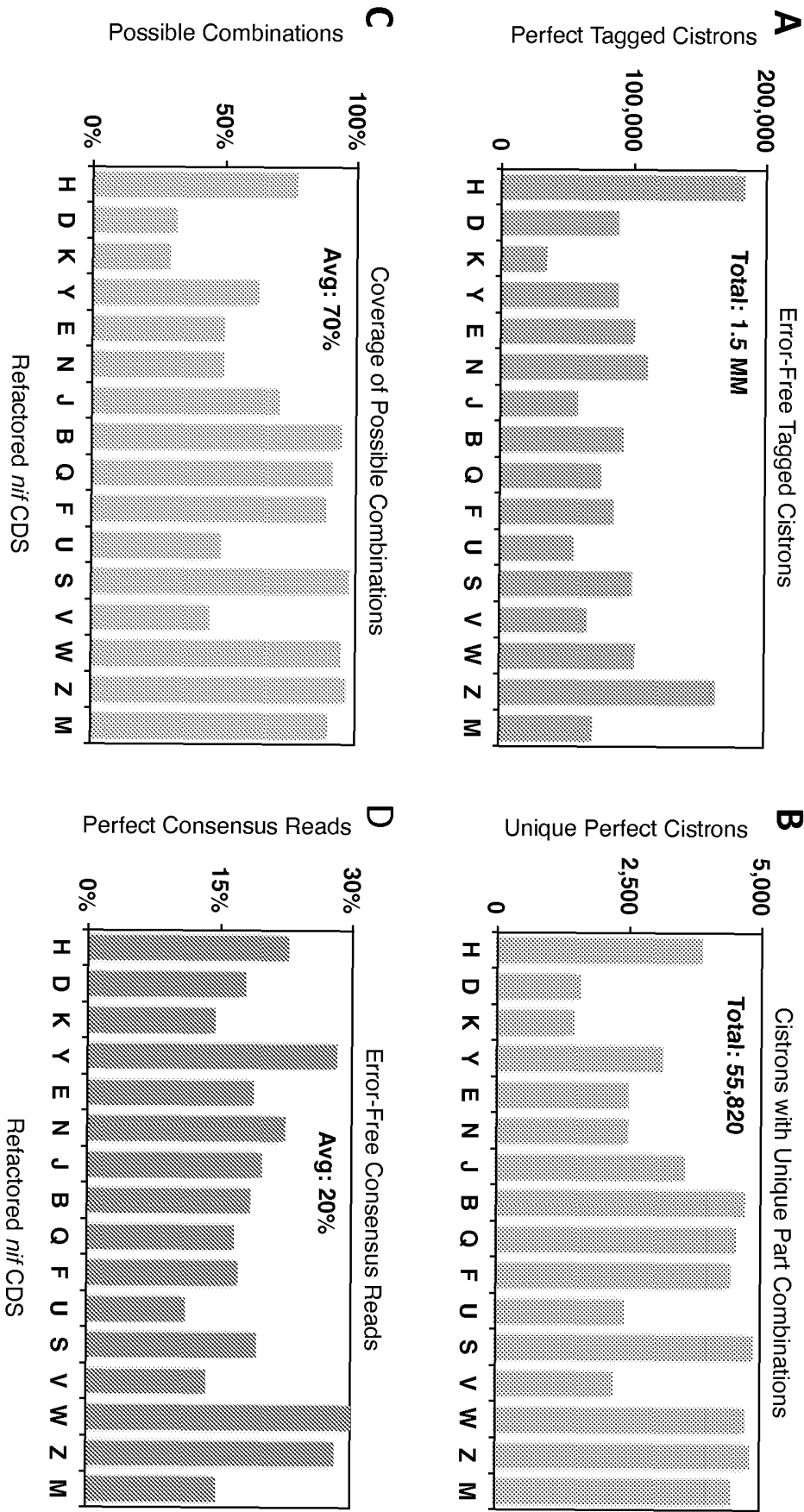


FIGURE 13

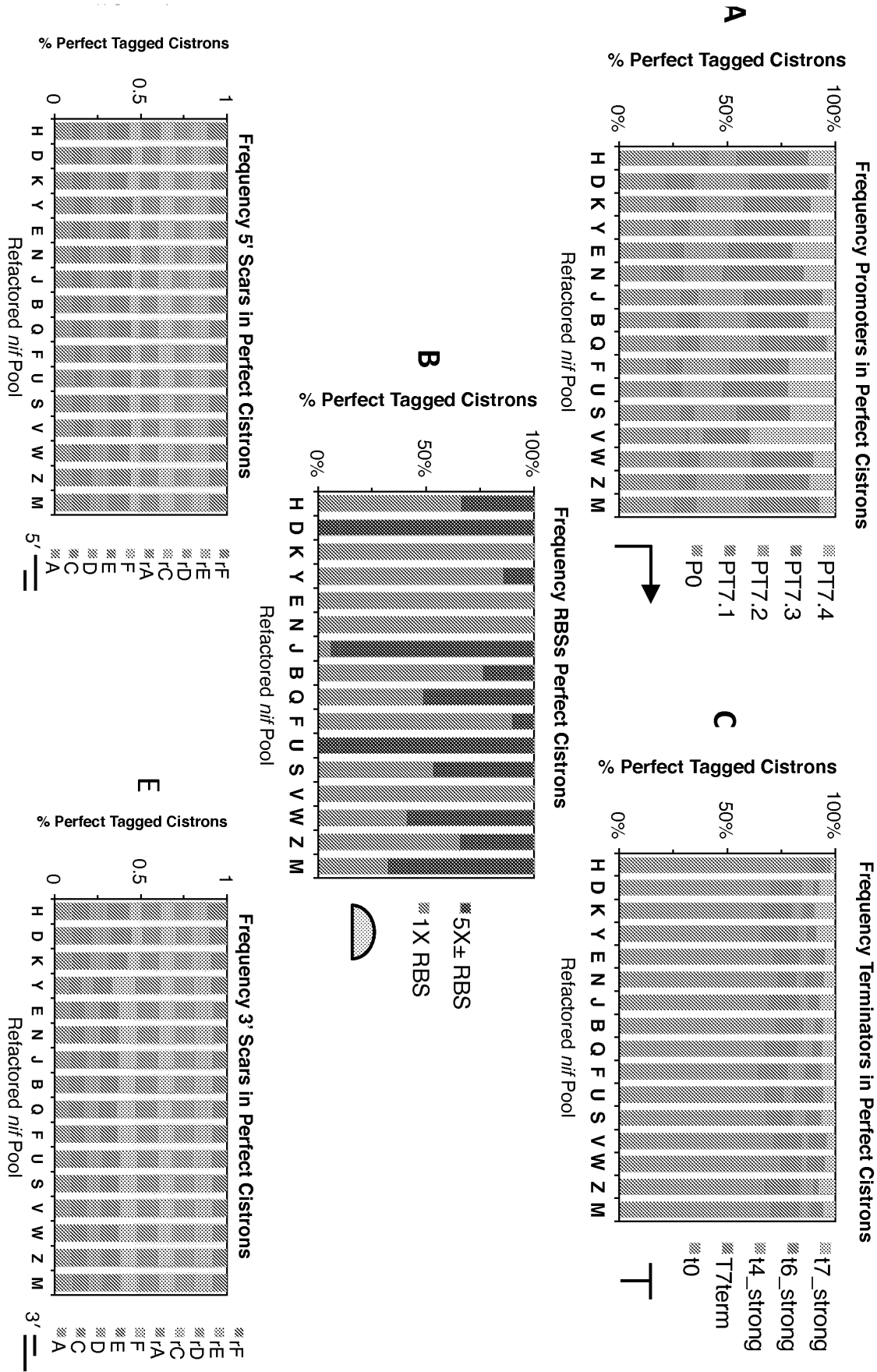


FIGURE 14

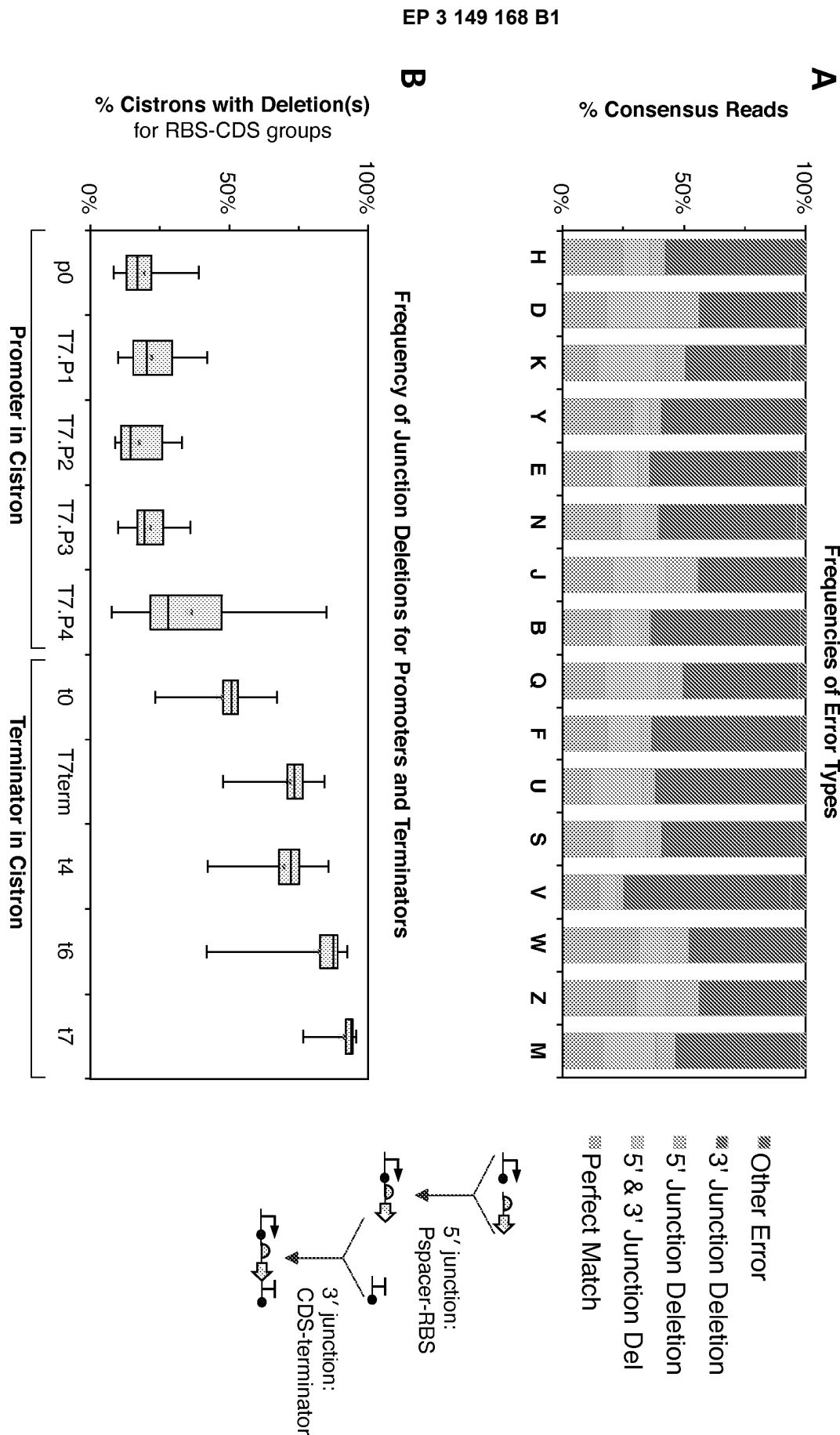


FIGURE 15

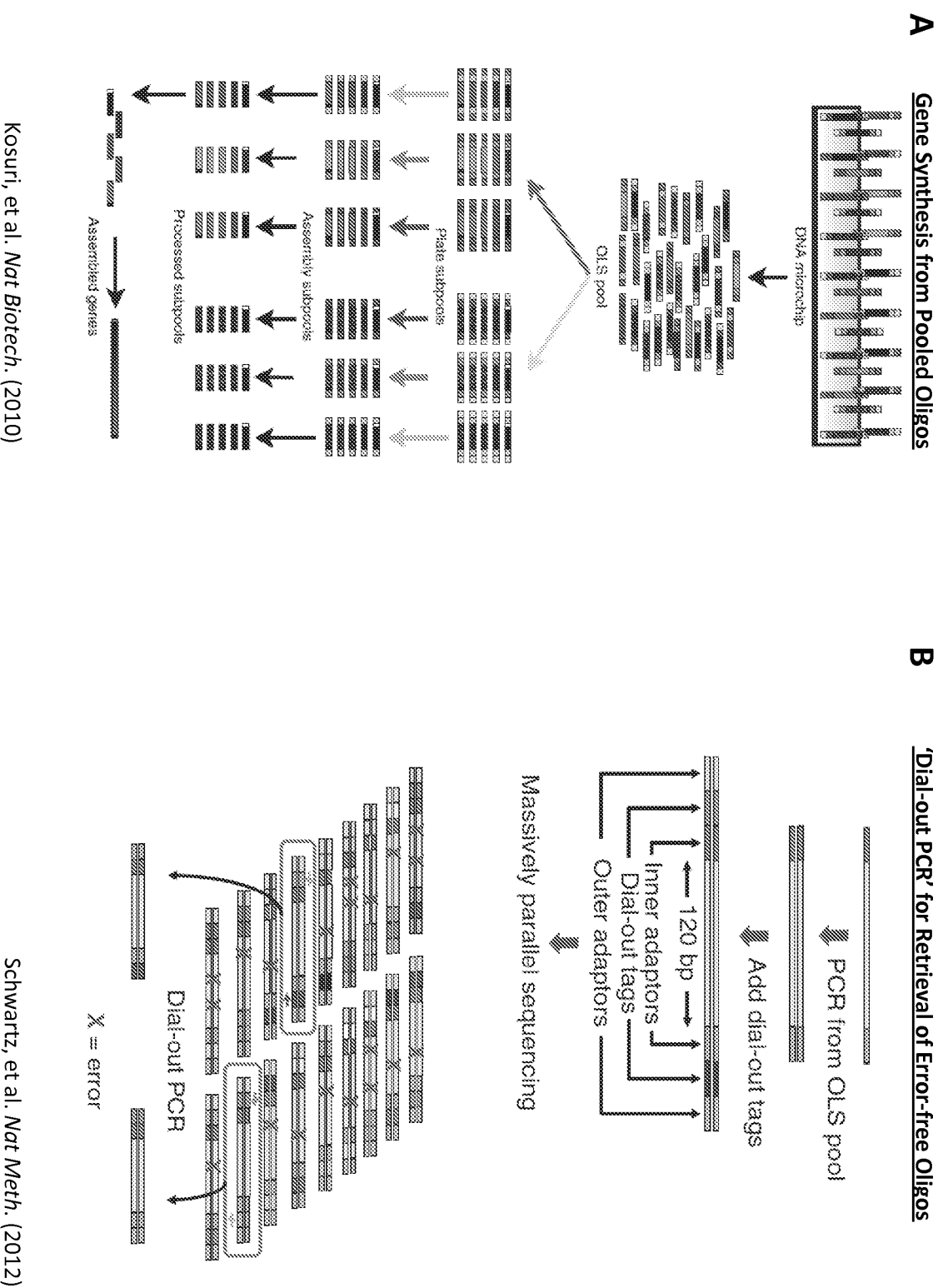


FIGURE 16

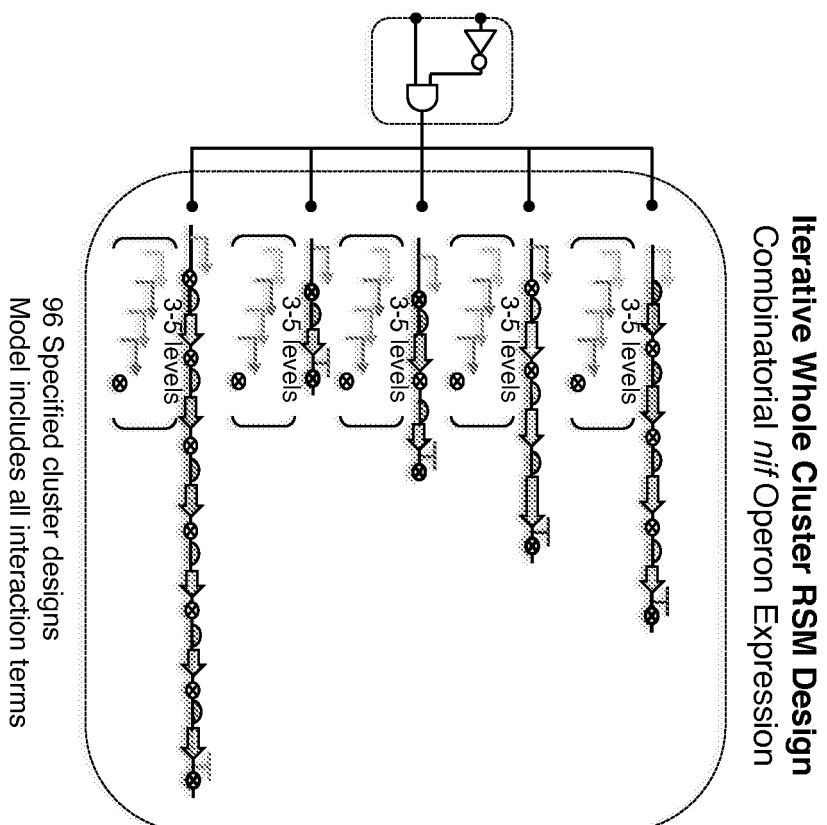


FIGURE 17

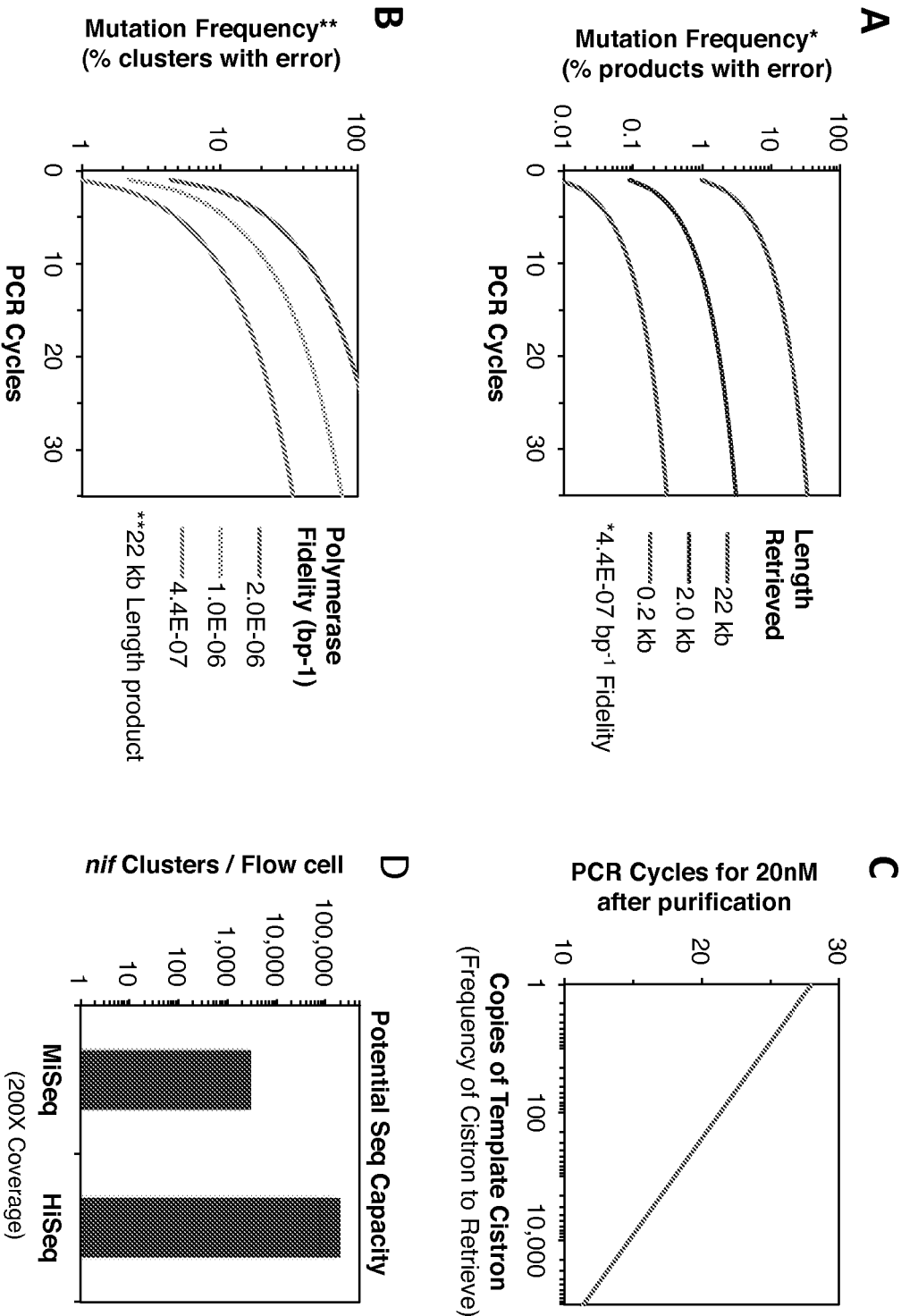


FIGURE 18

1. Multiplex 'Jumping' Libraries
 - Any length designs
 - In development
 - Universal and NGS compatible
2. Primer Walking Amplification
 - $\leq 2\text{kb}$ designs
 - Requires fixed positions/linkers
 - Relatively easy for cistrons
 - Not universally applicable
3. Single-Molecule Long Reads (SMRT)
 - $\leq 5\text{-}8\text{kb}$ designs
 - Universal if 2° structure ok
 - Will test on cistron pools

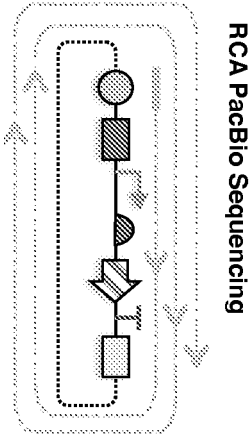
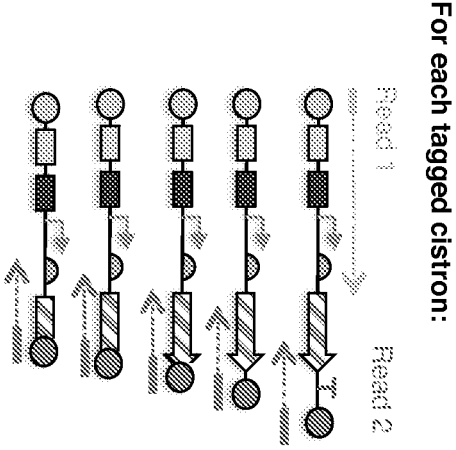


FIGURE 20

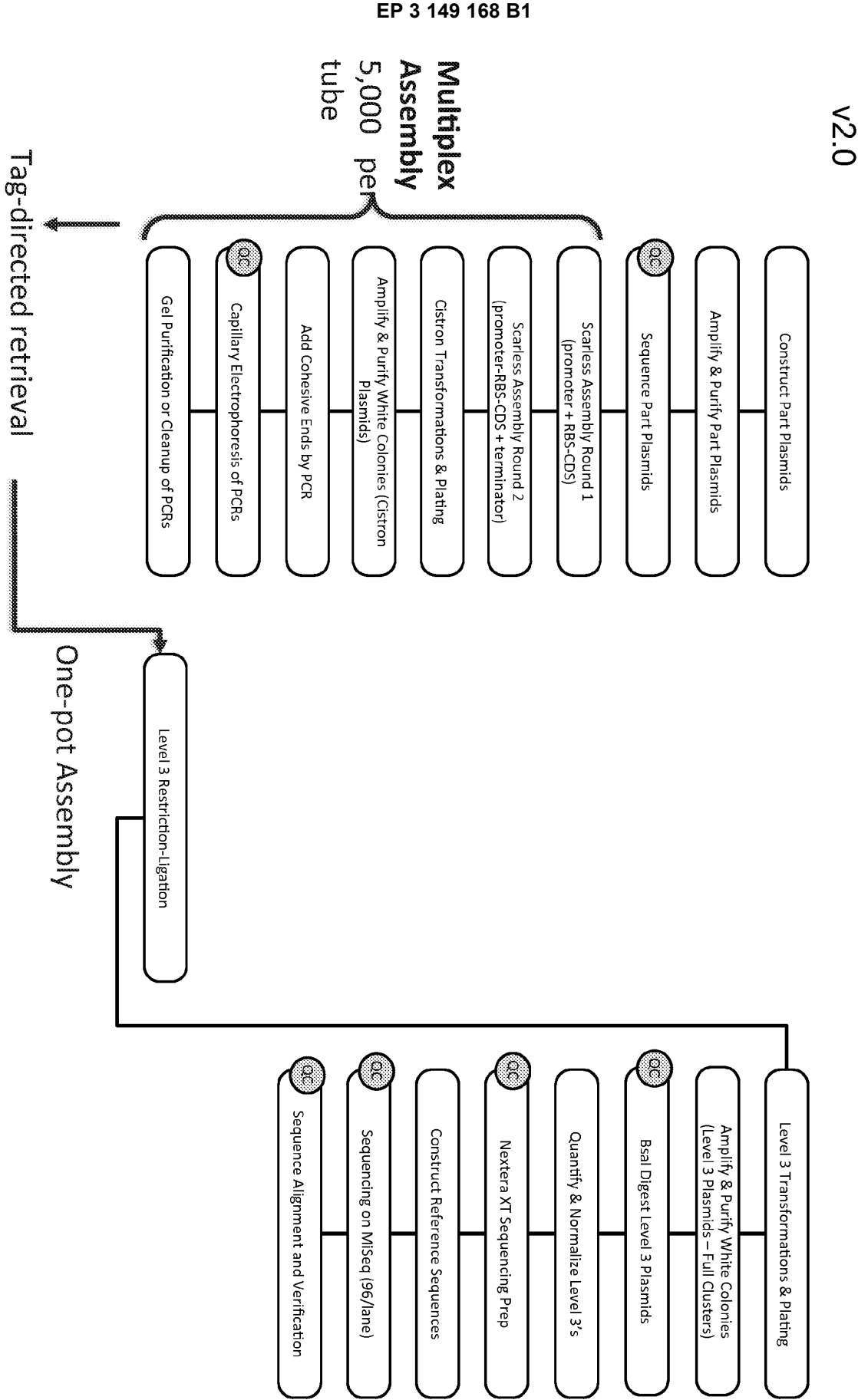


FIGURE 21

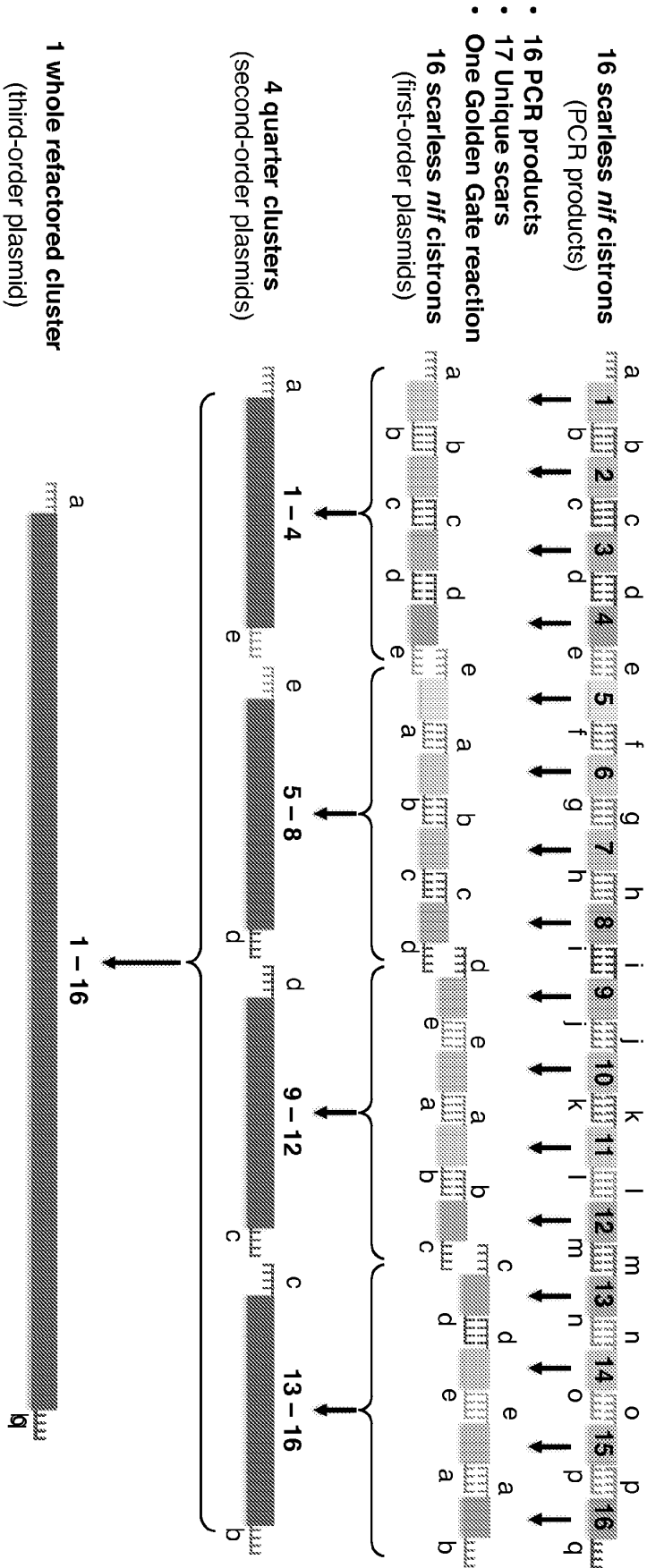
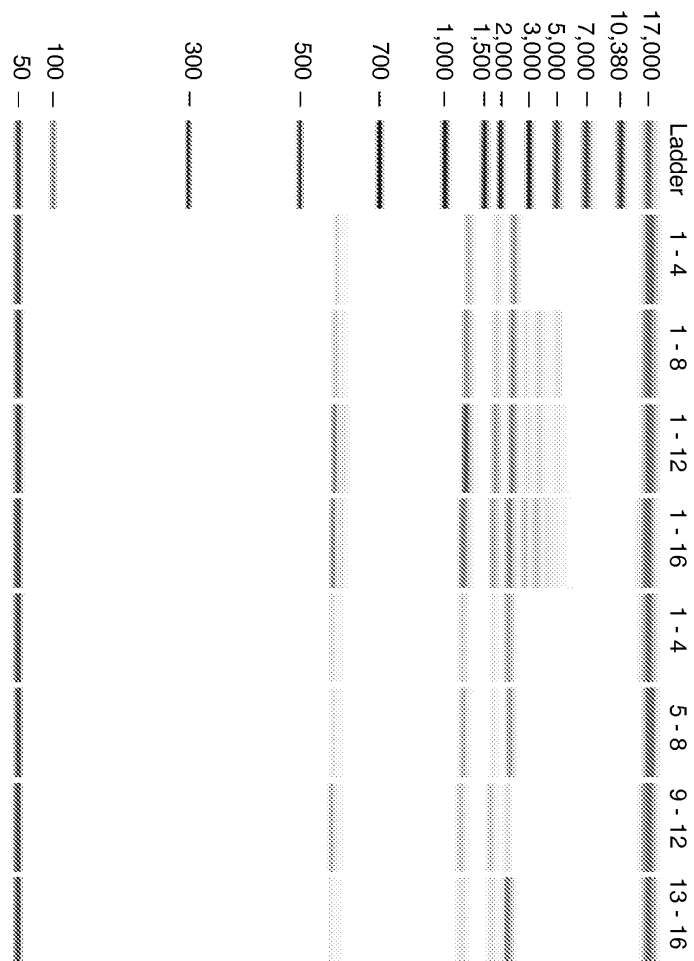


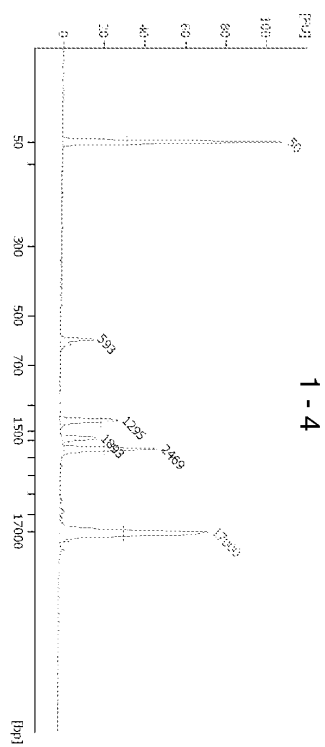
FIGURE 22

A

MoClo conditions (Weber
et al, 2011)

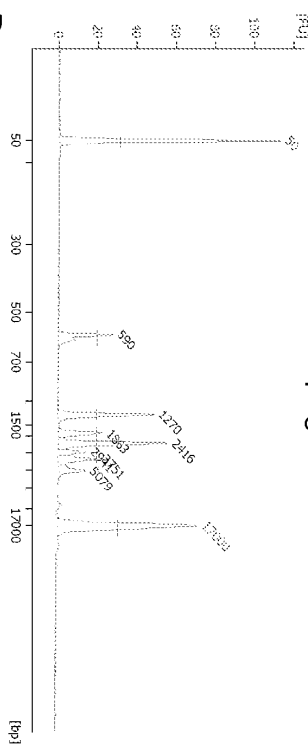


B



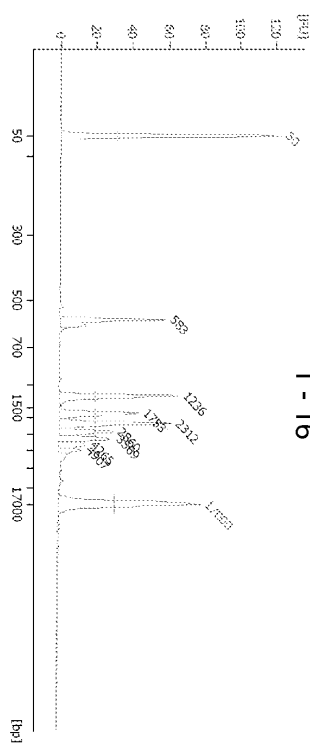
1 - 4

C



1 - 8

D



1 - 16

FIGURE 23

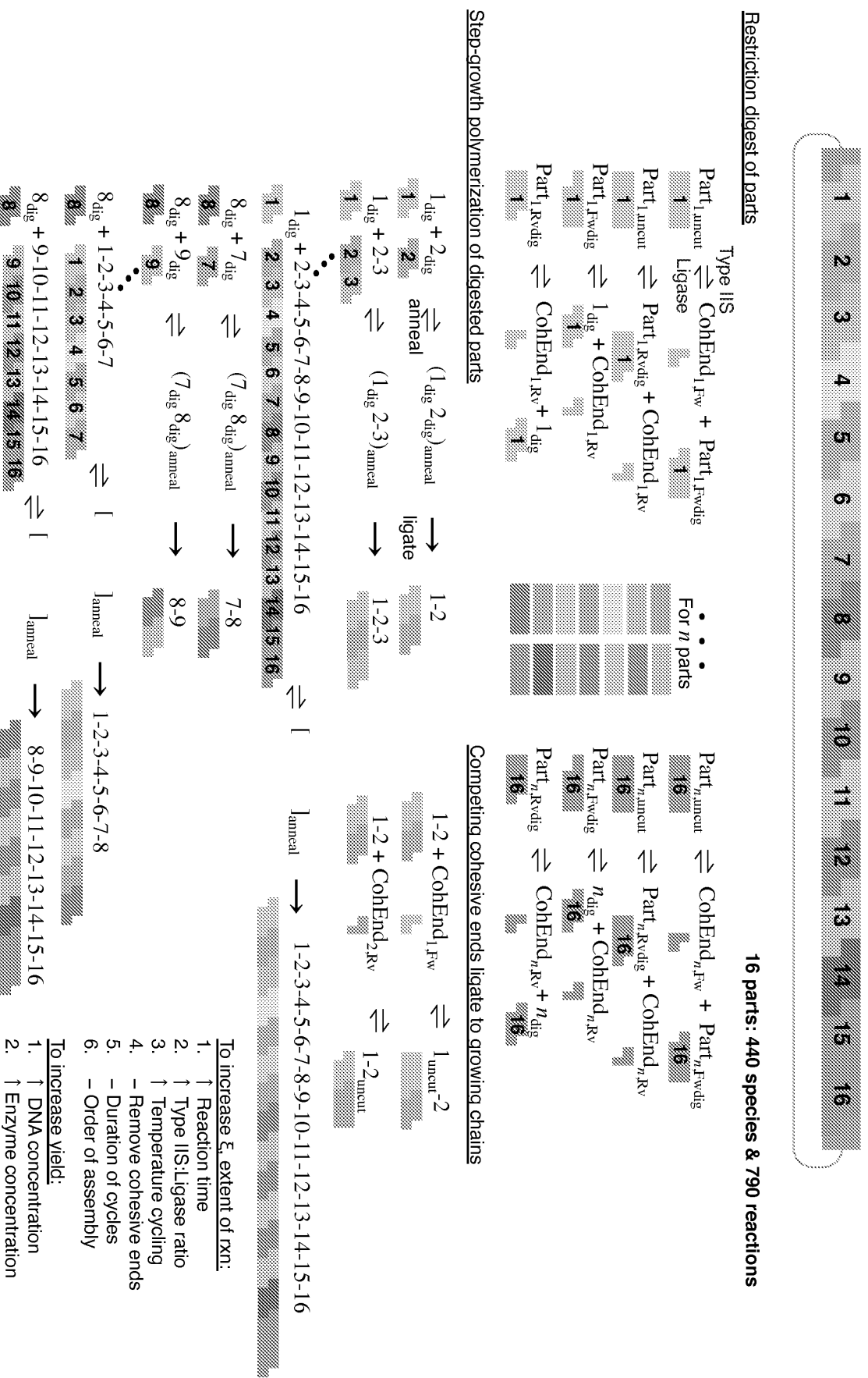


FIGURE 24

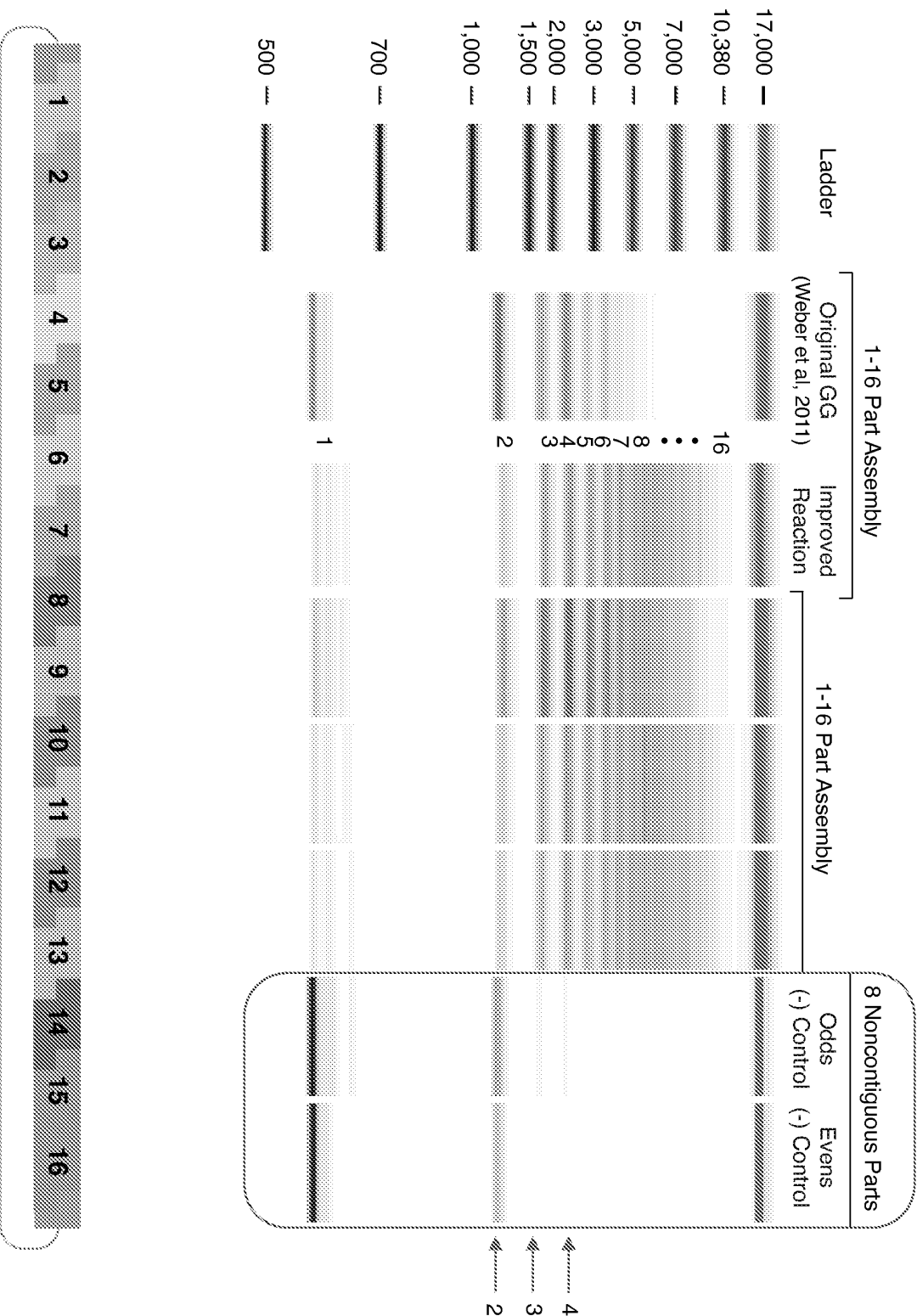


FIGURE 25

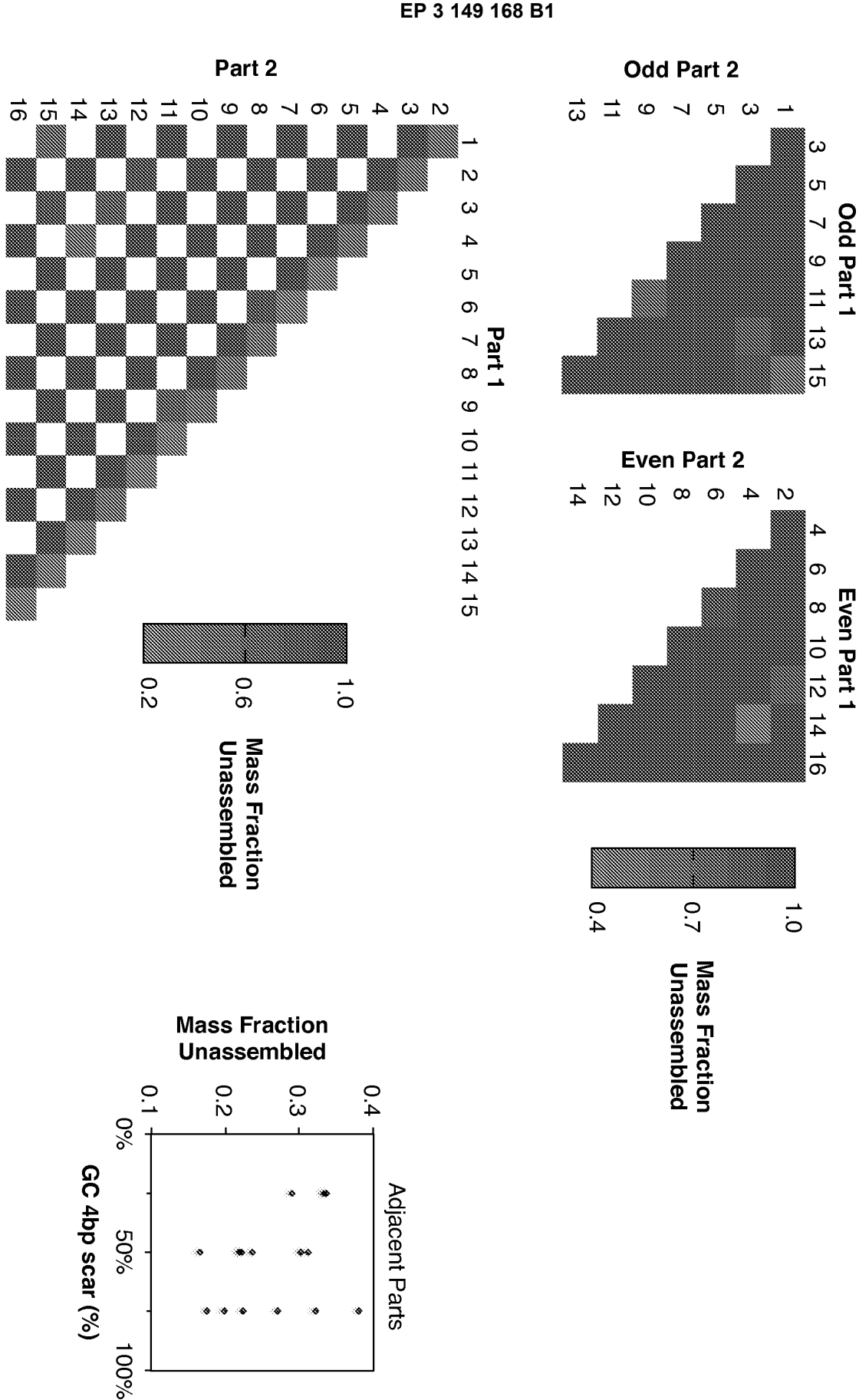
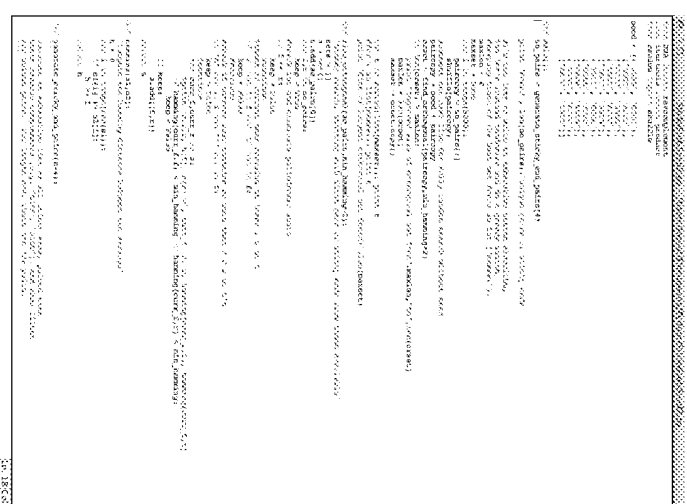
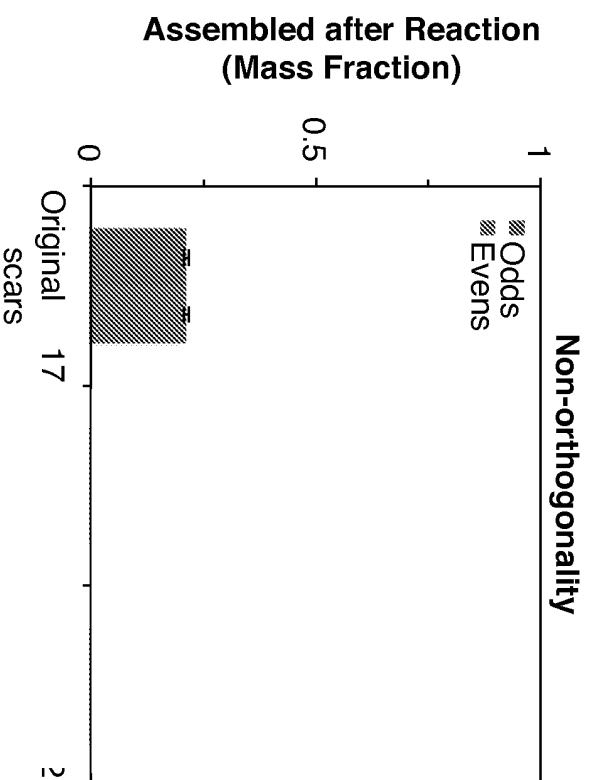


FIGURE 26

Design Rules for Cohesive Ends

1. Cannot share >2 nt with any scar
2. Cannot share >2 nt with RC of any scar or itself
3. Must contain 1 G or C
4. Cannot contain > 3 G's or C's
5. Must be 4 nt in length



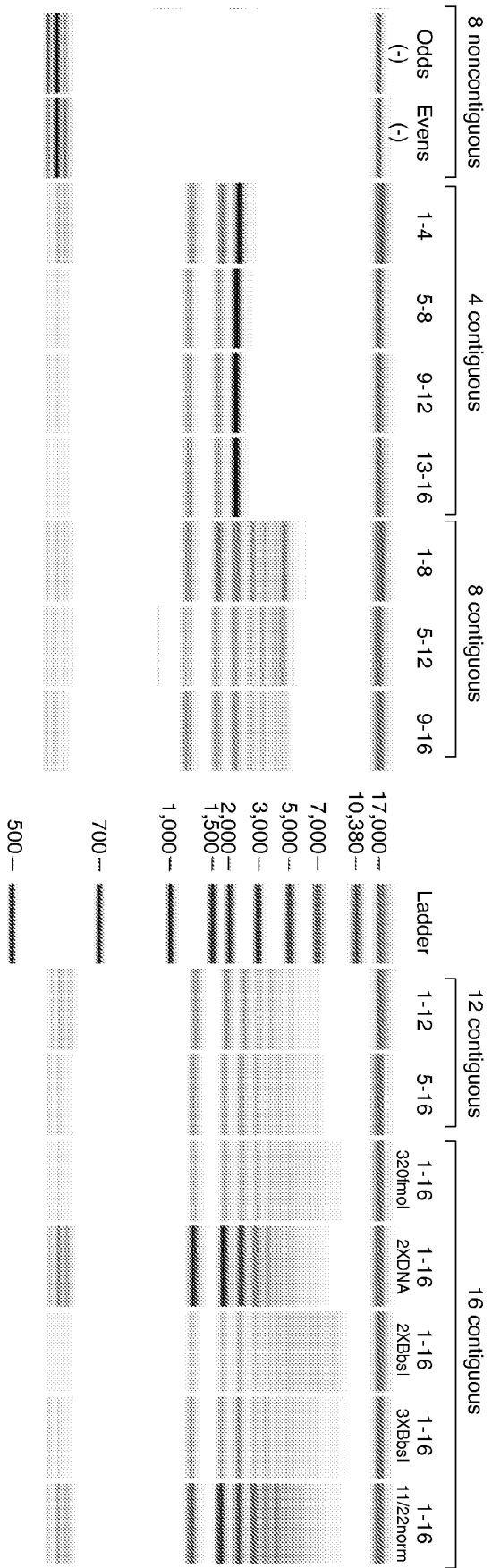


FIGURE 27

FIGURE 28

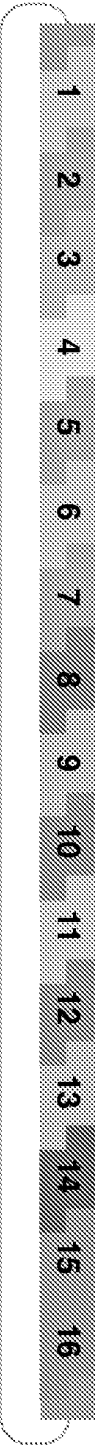
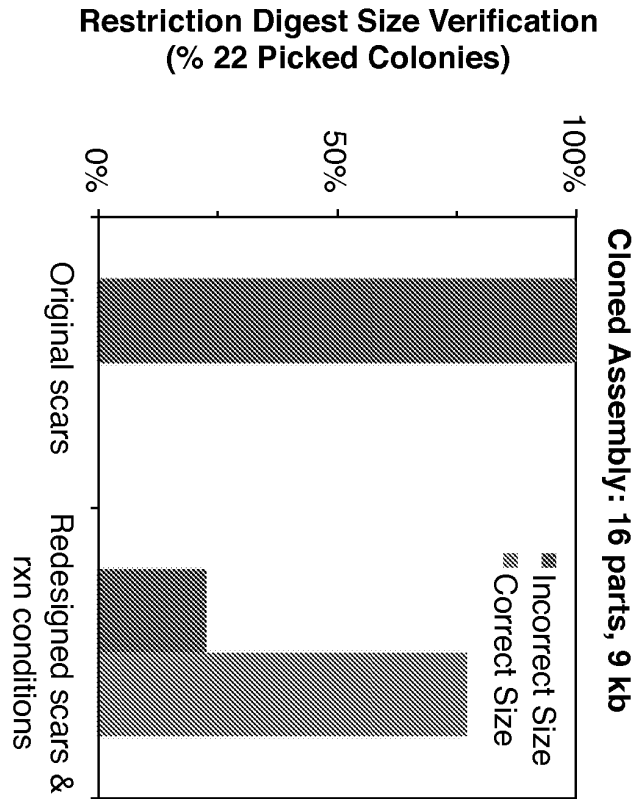


FIGURE 29

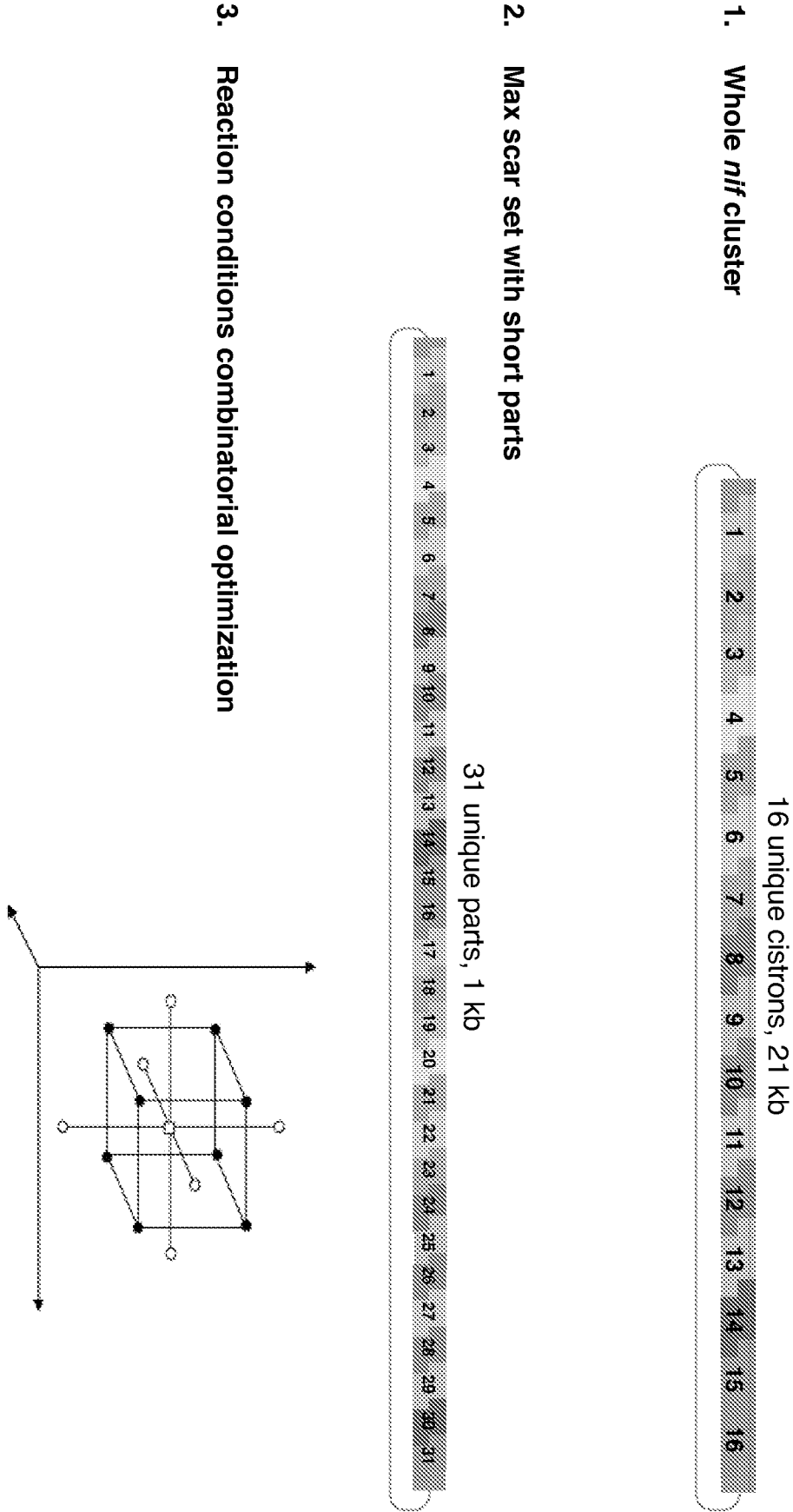


FIGURE 30

Accessing complex biological functions

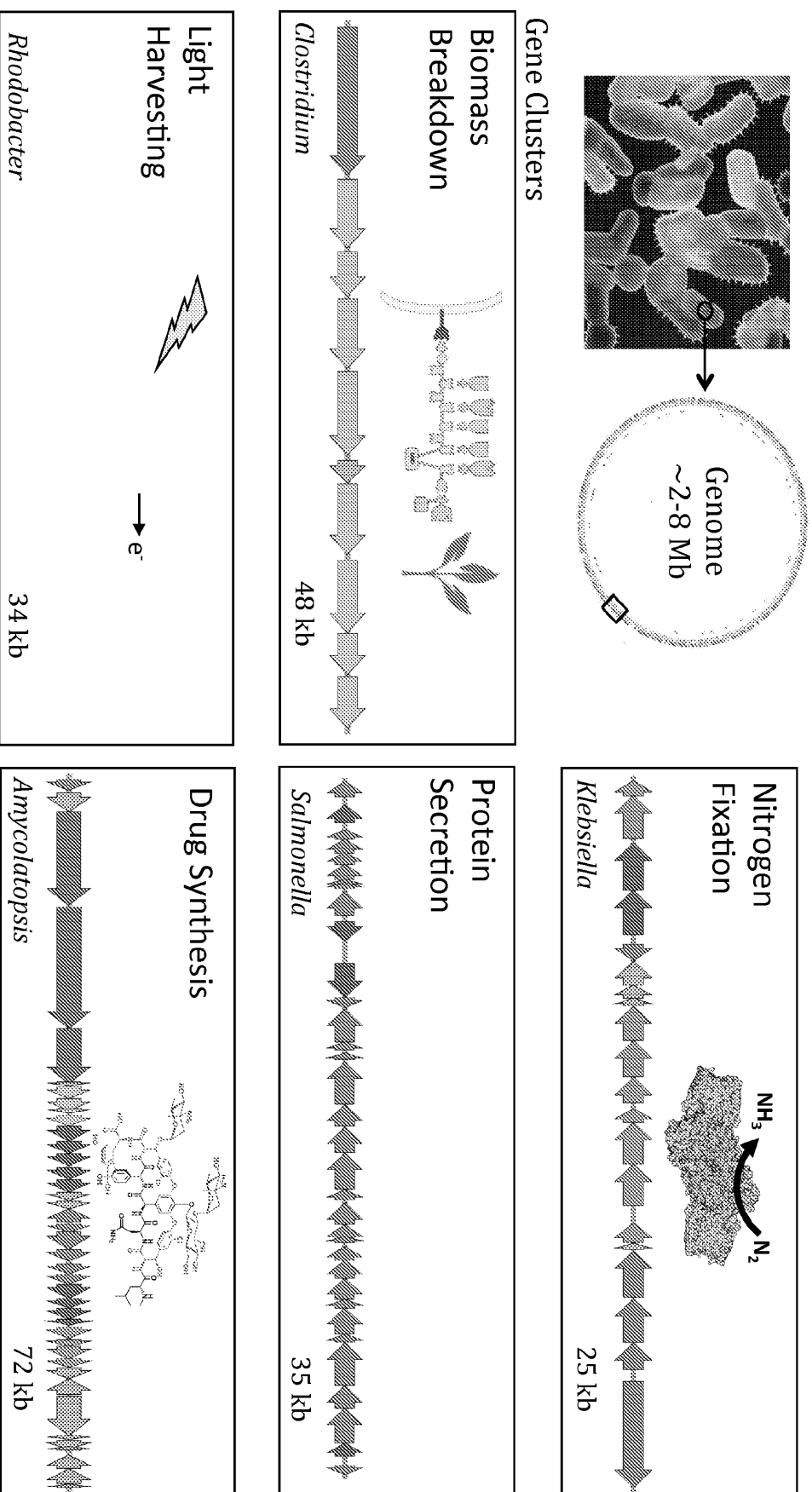


FIGURE 31

Refactoring Gene Clusters for Control and Modularity

Precision Gene Expression Engineering

- Transcriptional control
- Translational control
- Genetic insulation

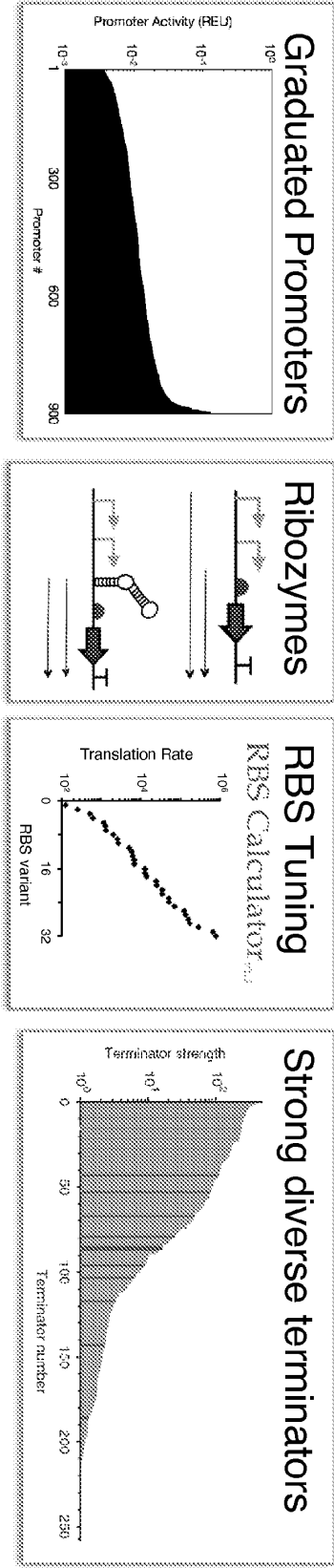
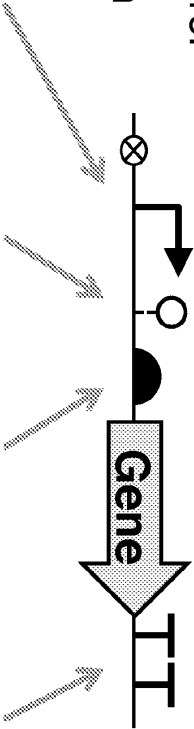


FIGURE 32

Building permuted clusters hierarchically from parts

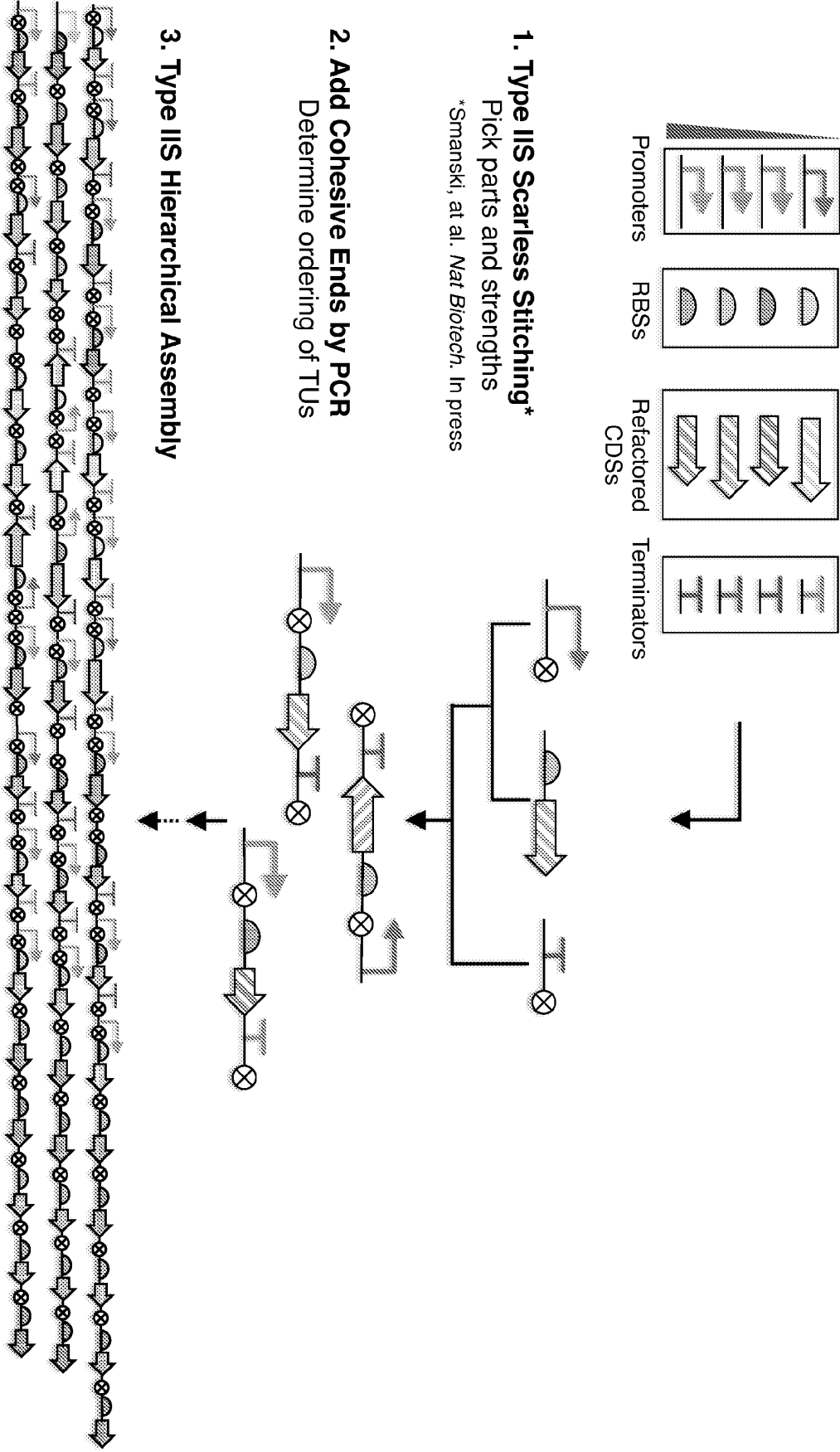


FIGURE 33

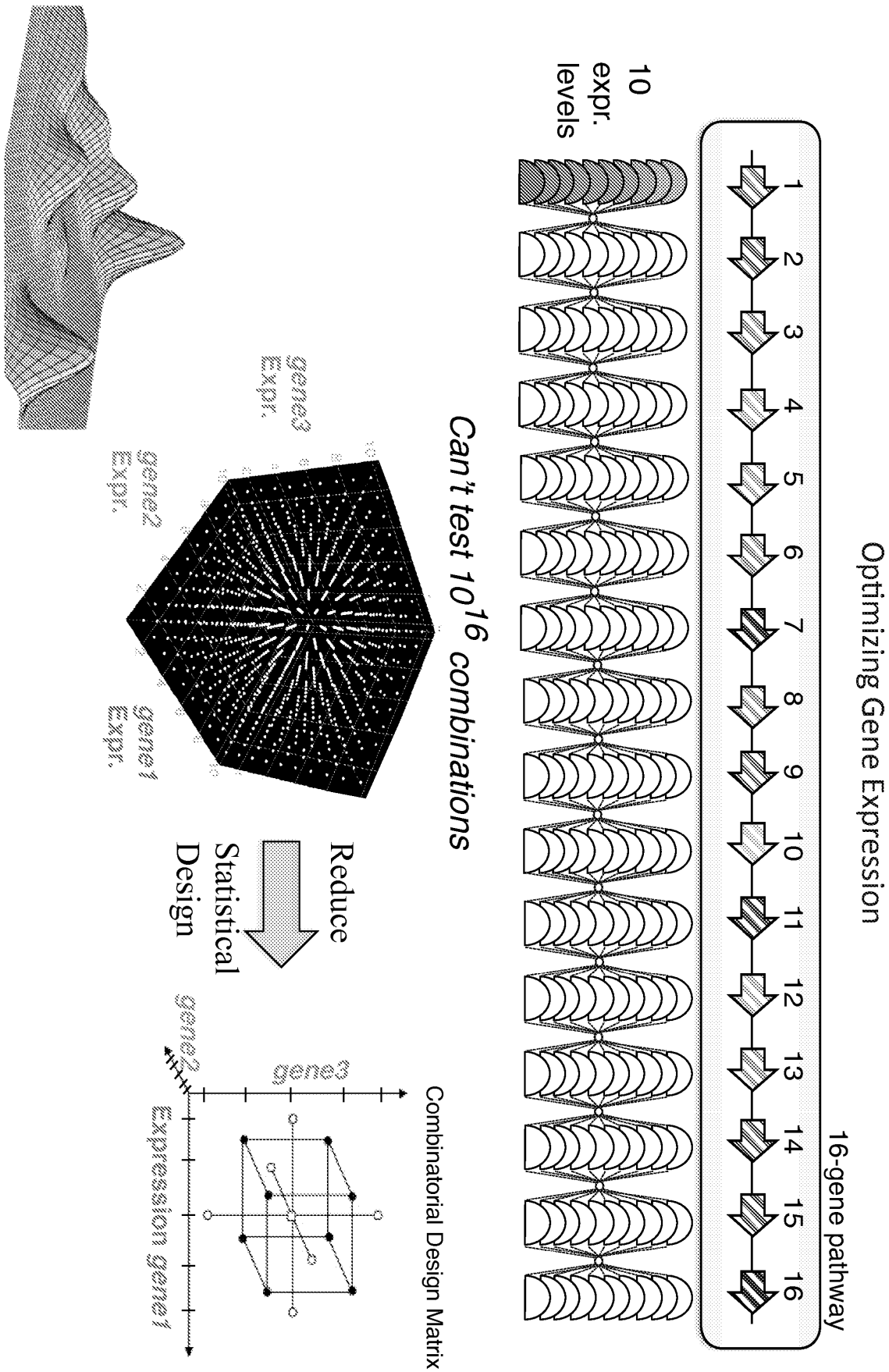


FIGURE 34

Streamlined construction enables larger, more complex libraries

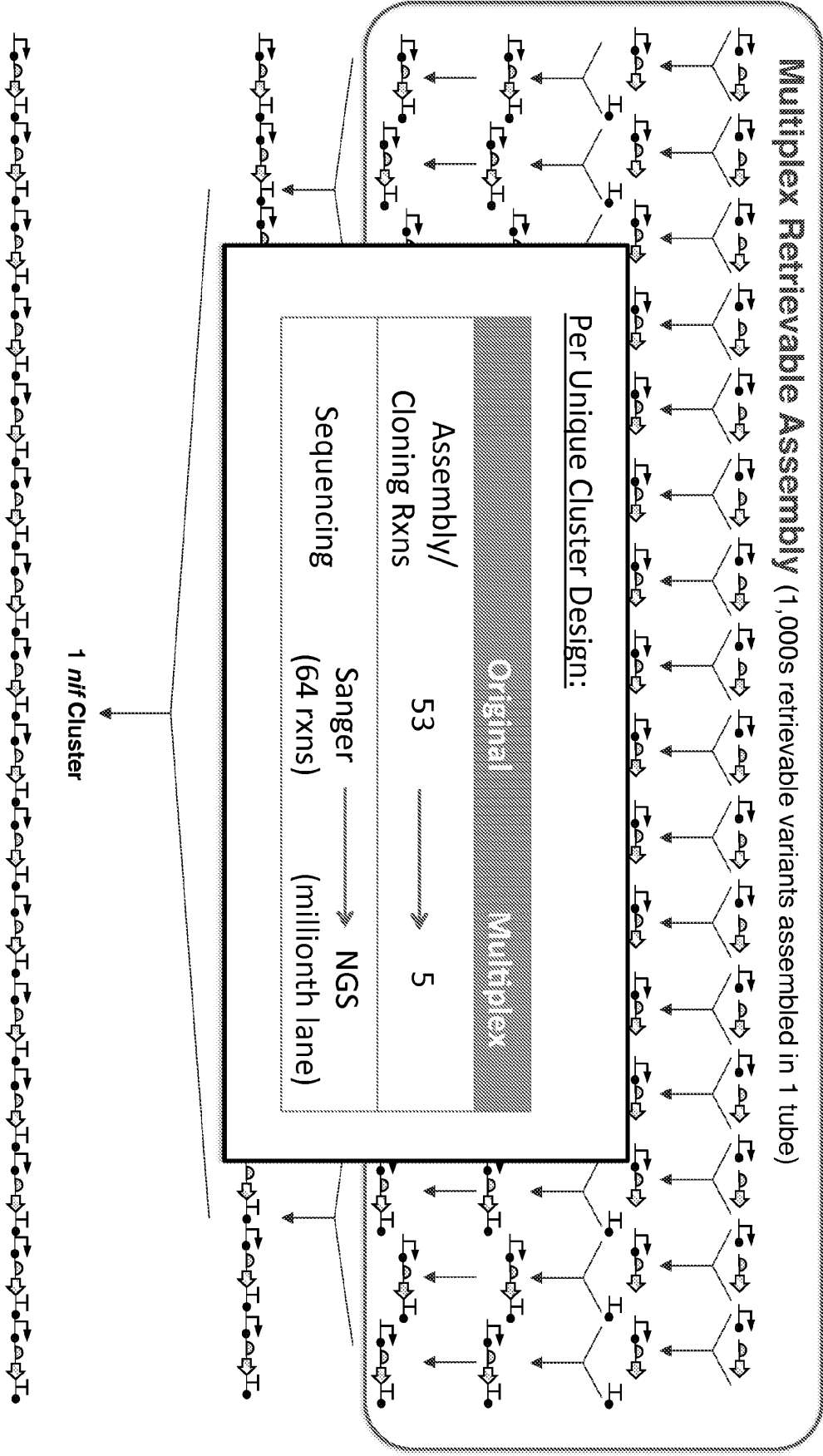


FIGURE 35
>100-fold increase in library size for N₂ fixation cluster

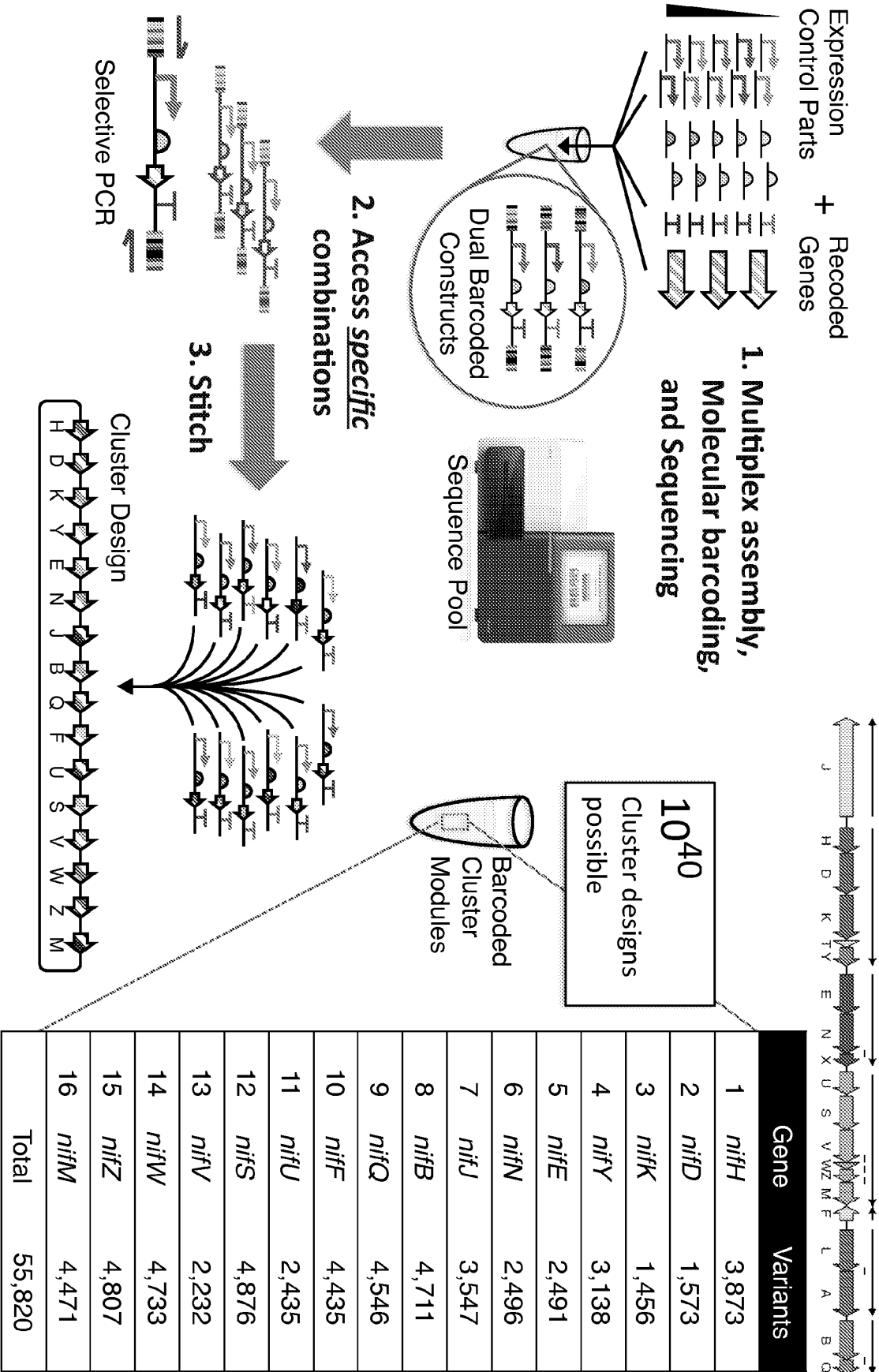


FIGURE 36
7.5M cistrons sequenced identify error modes

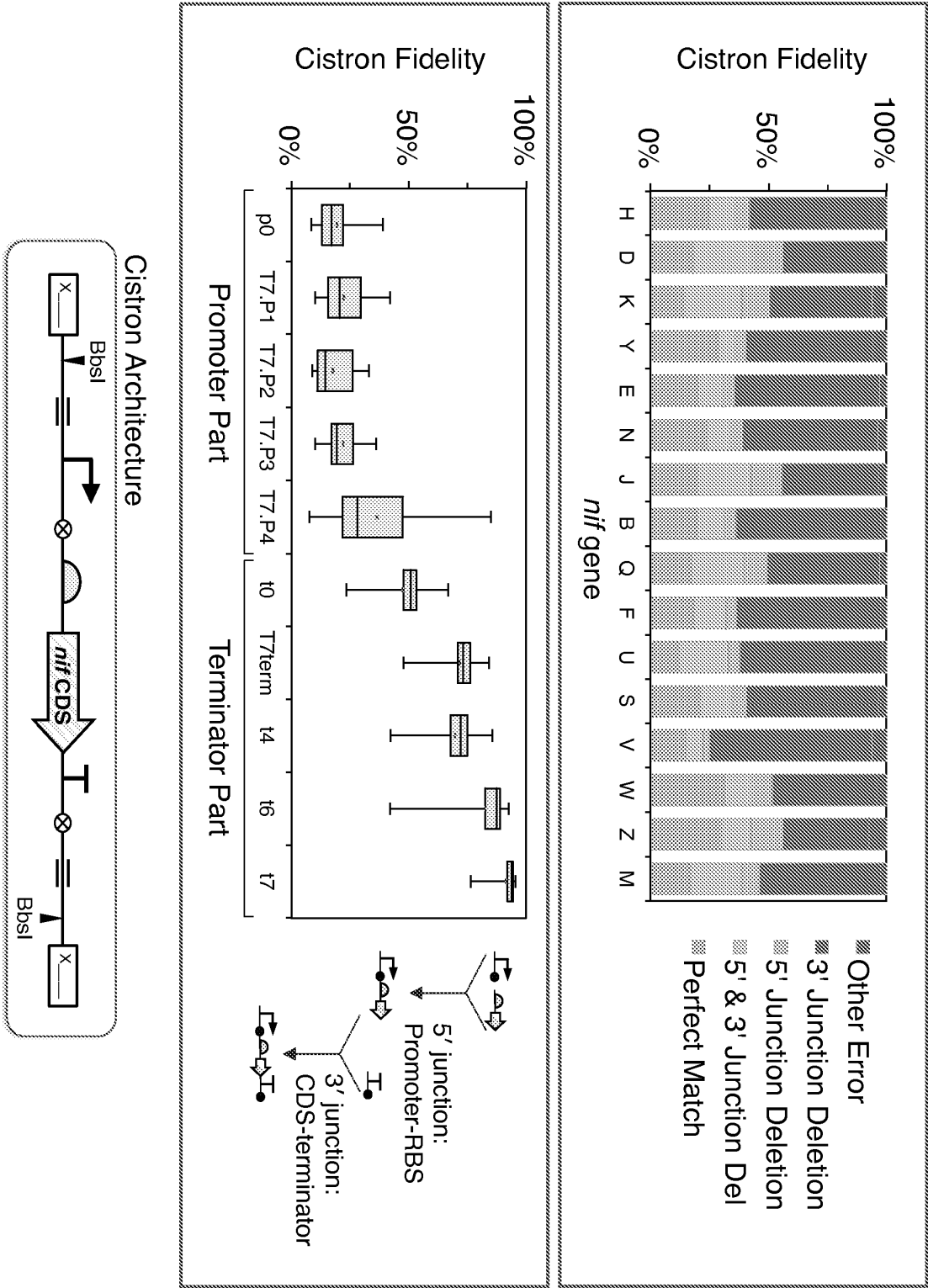


FIGURE 37

Rapid iterative permuting of multi-gene pathway

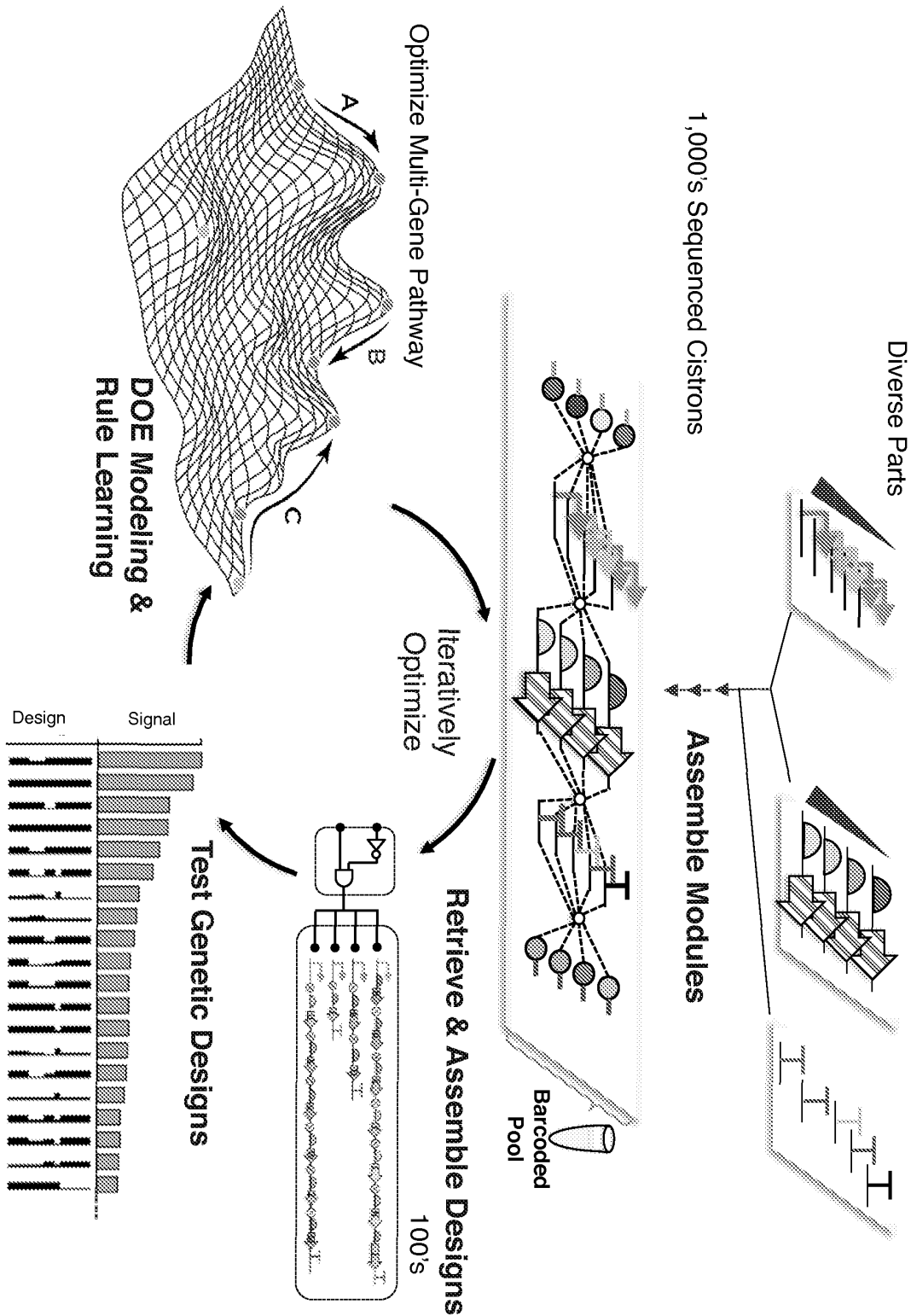


FIGURE 38
>95% Retrieved cistrons error-free (84 kb)

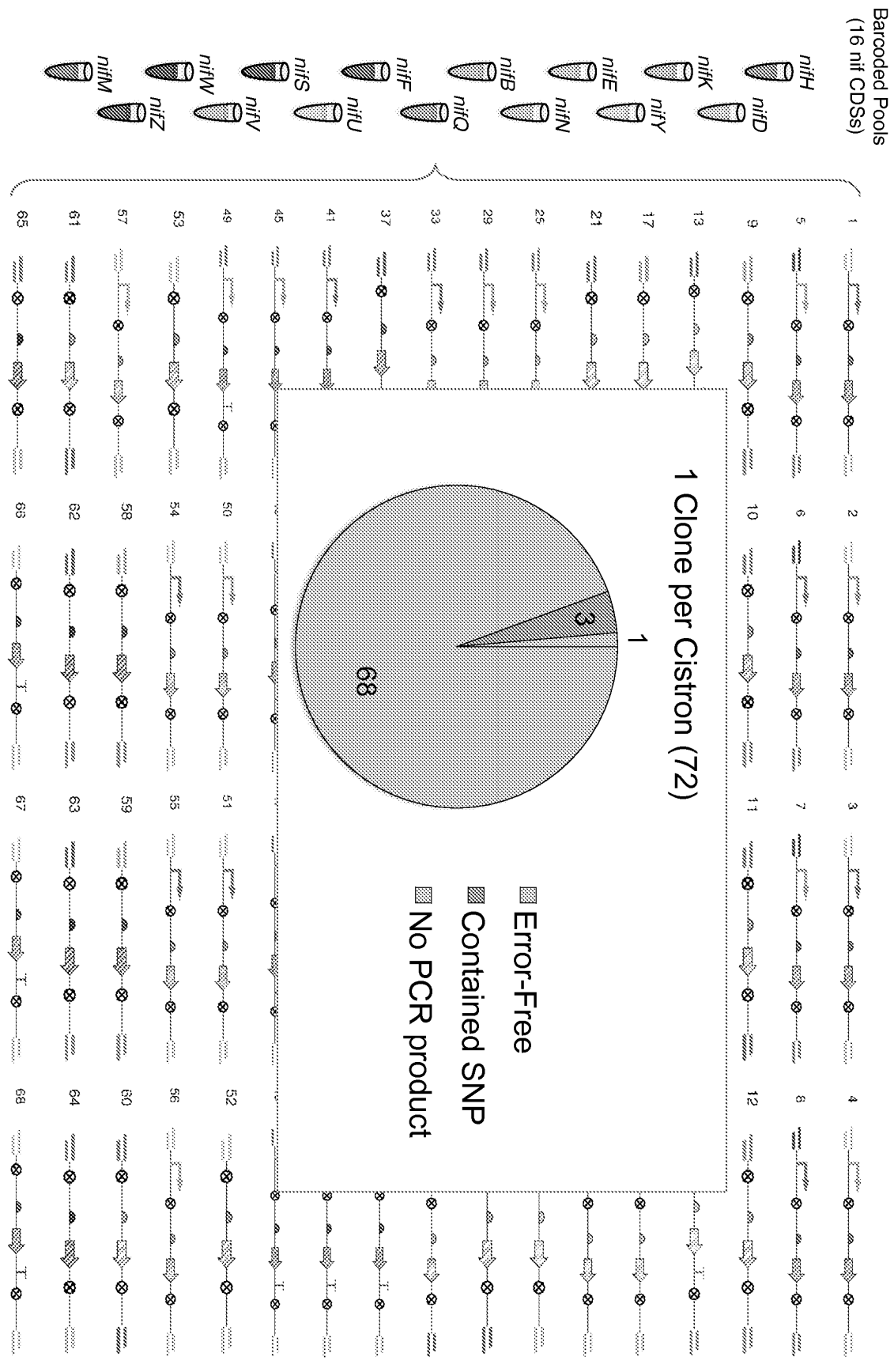
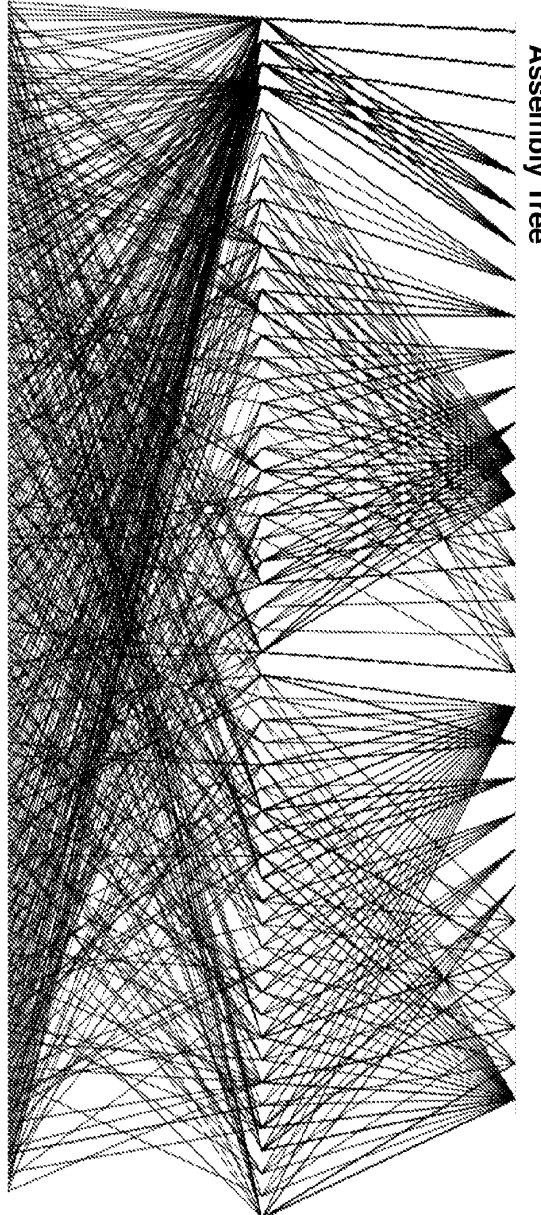
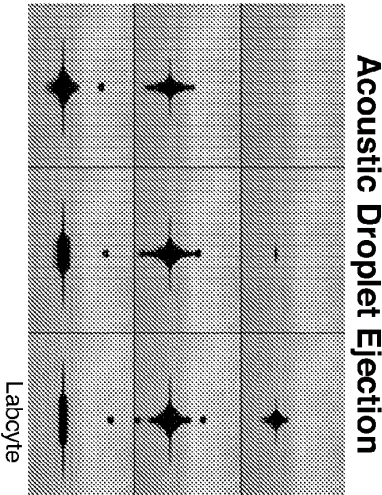
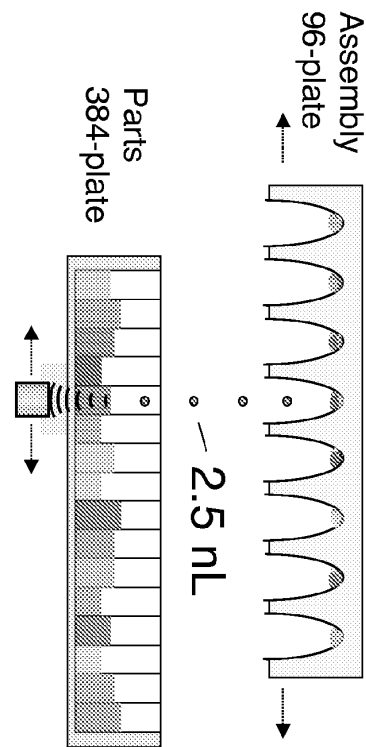


FIGURE 39

Automated and Miniaturized Assembly



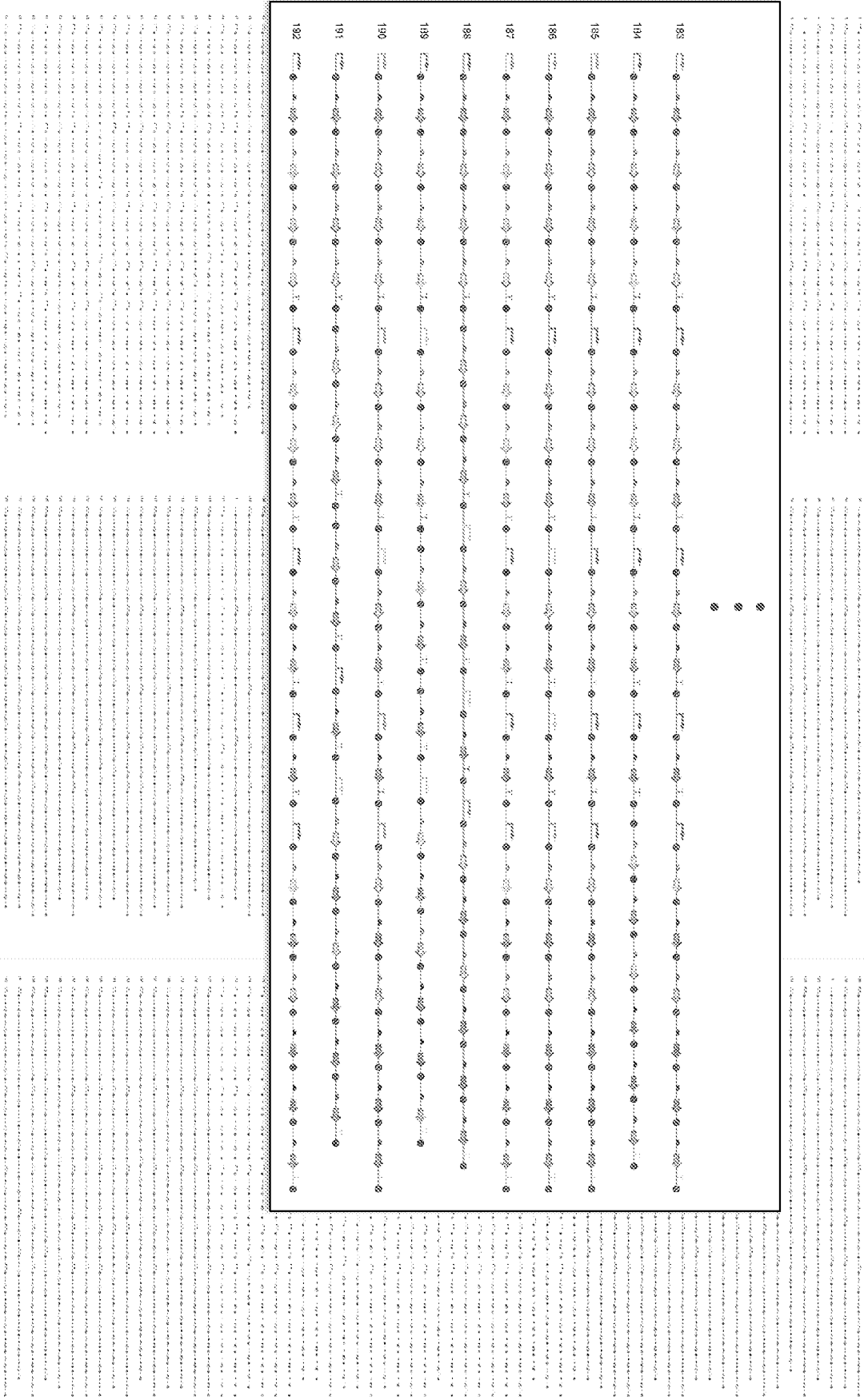
35 Cistrons

55 Quarter clusters

192 nif Clusters (4.5 Mb)
7.7% of 2,500 possible

FIGURE 40

192 Systematically permuted *nif* clusters (4.5 Mb)



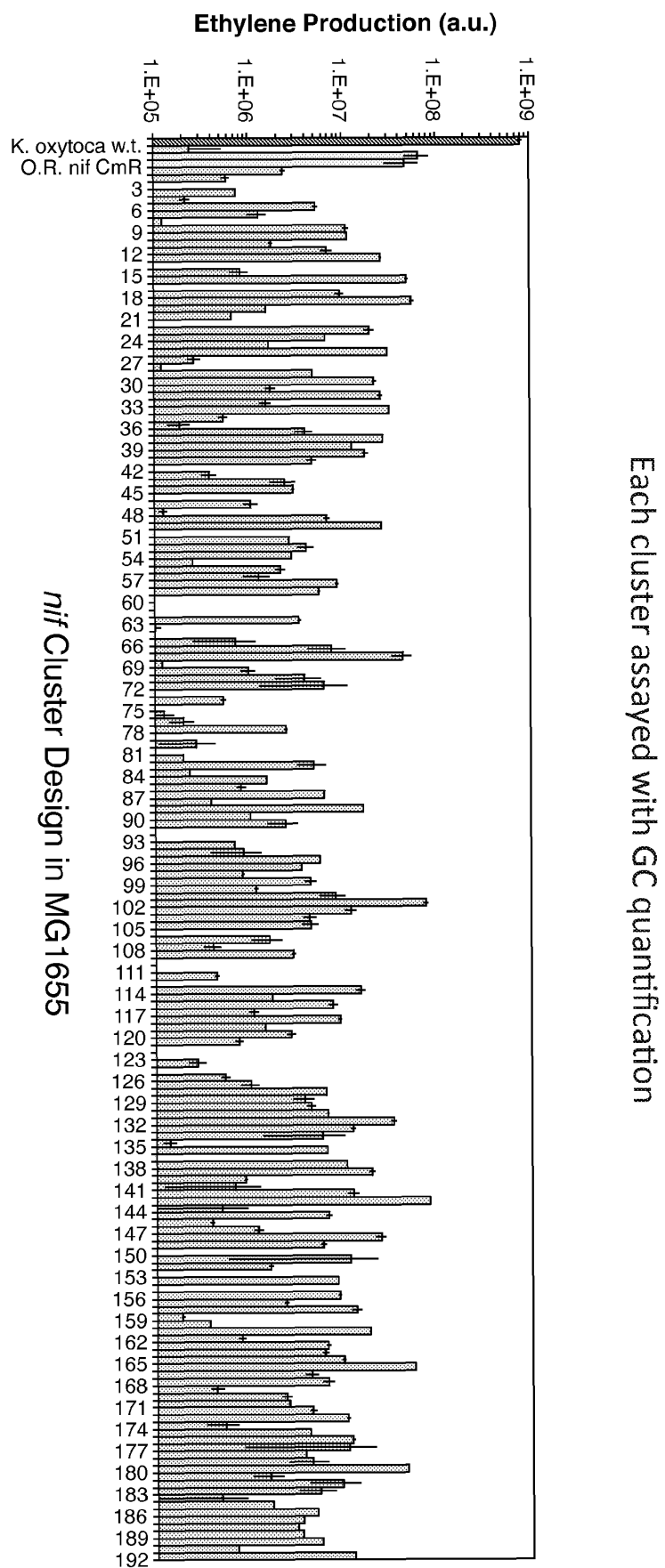
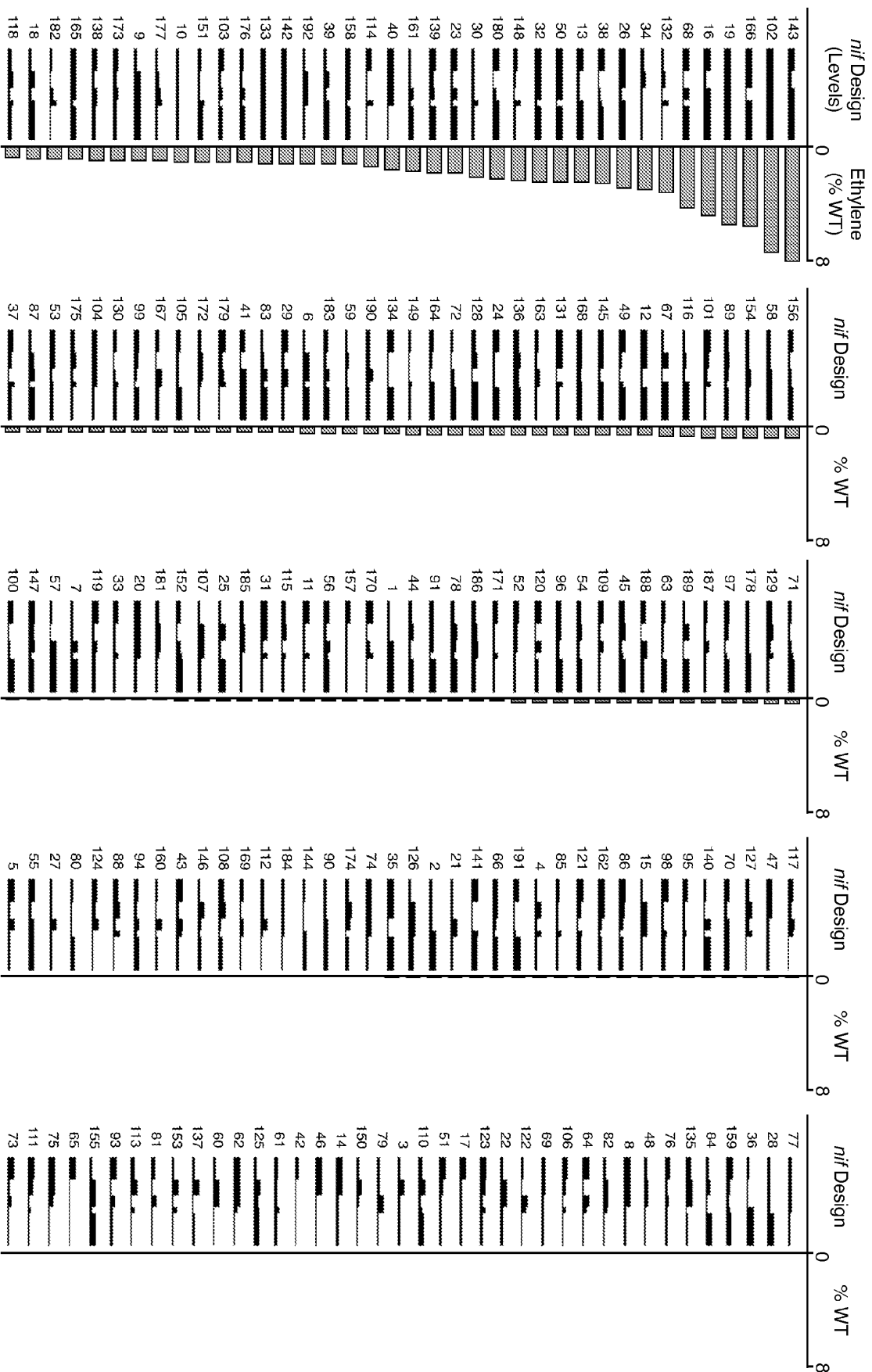


FIGURE 41

FIGURE 42
Mapped cluster activity to genetic designs



Multivariate modeling identifies main effects and interactions

Actual vs. Predicted Activity 2-level factorial

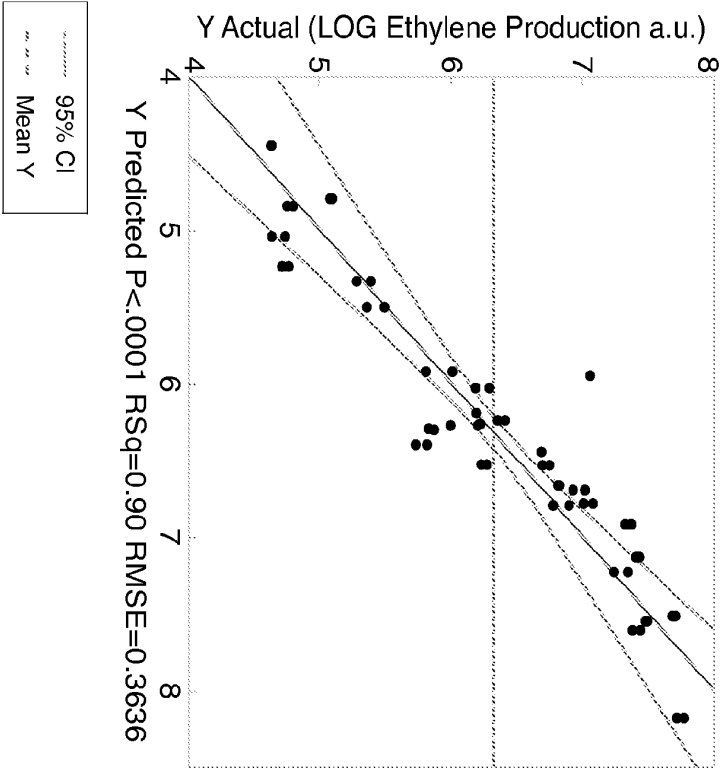
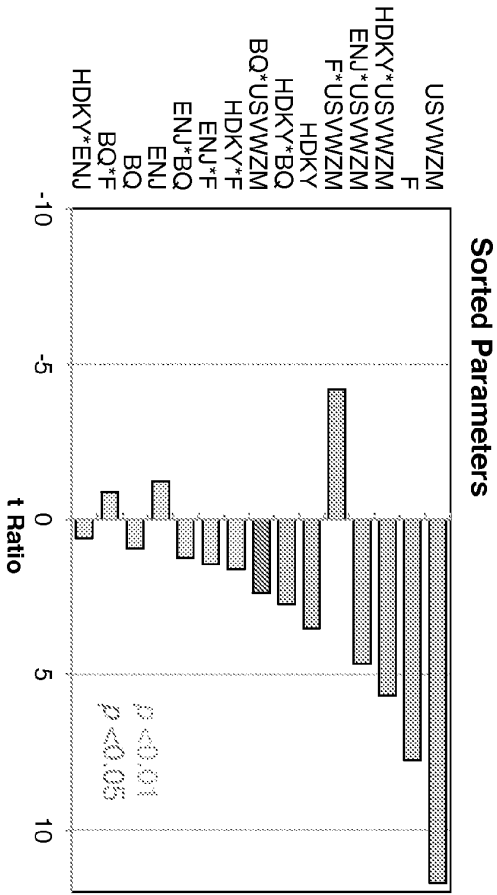


FIGURE 43



Term	Estimate	Std Error	t Ratio	Prob> t
USVWZM	0.573	0.049	11.71	<.0001*
F	0.386	0.050	7.76	<.0001*
HDKY*USVWZM	0.282	0.050	5.67	<.0001*
ENJ*USVWZM	0.229	0.049	4.66	<.0001*
F*USVWZM	-0.206	0.049	-4.18	0.0001*
HDKY	0.171	0.049	3.49	0.0012*
HDKY*BQ	0.136	0.050	2.72	0.0096*
BQ*USVWZM	0.117	0.050	2.35	0.0234*
HDKY*F	0.080	0.050	1.59	0.121
ENJ*F	0.072	0.050	1.43	0.161
ENJ*BQ	0.062	0.050	1.23	0.225
ENJ	-0.060	0.049	-1.22	0.229
BQ	0.046	0.050	0.93	0.356
BQ*F	-0.045	0.050	-0.90	0.376
HDKY*ENJ	0.030	0.050	0.60	0.555

FIGURE 44
Multidimensional response surface resolved by DOE

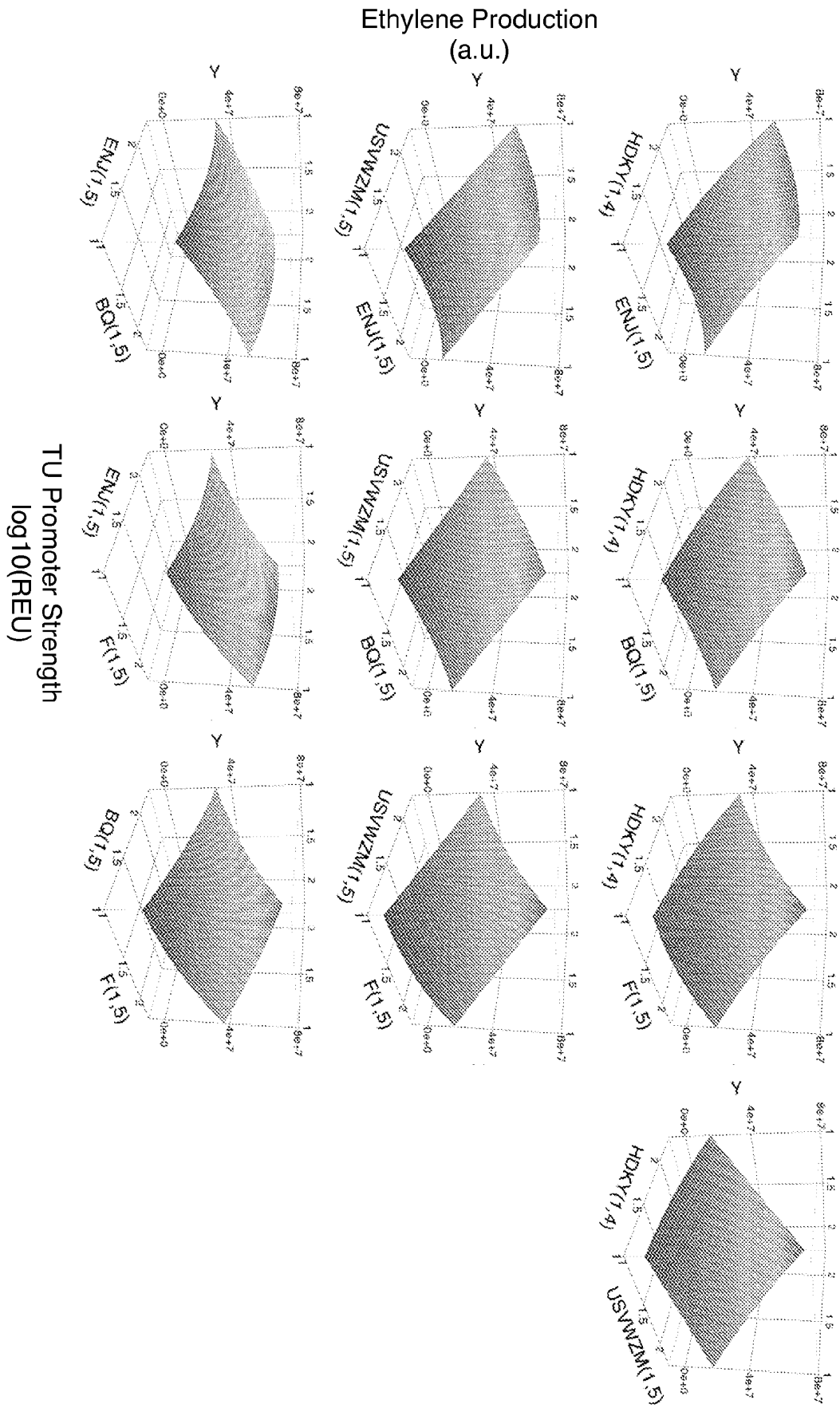


FIGURE 45
Model-guided redesign to test predictions

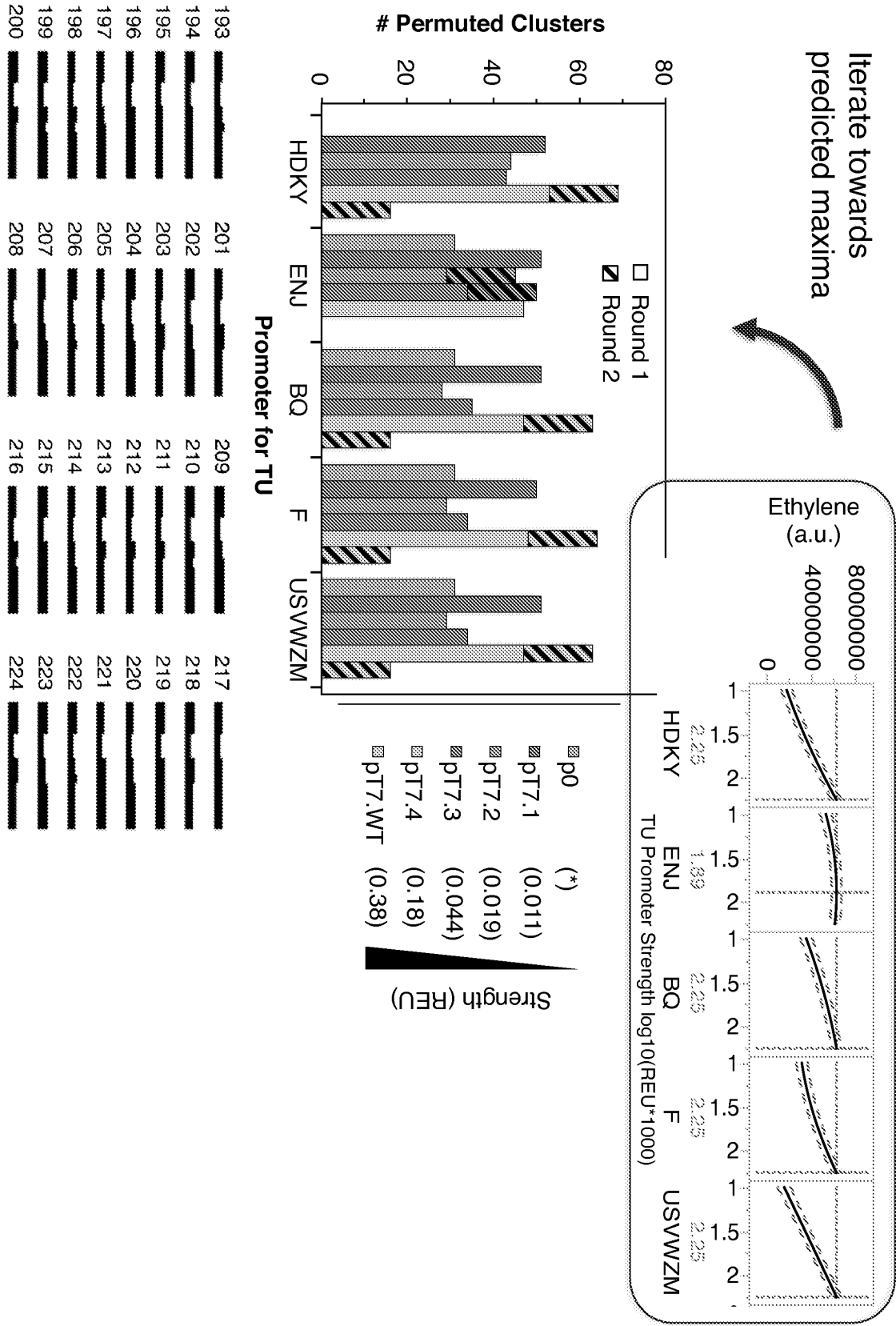


FIGURE 46

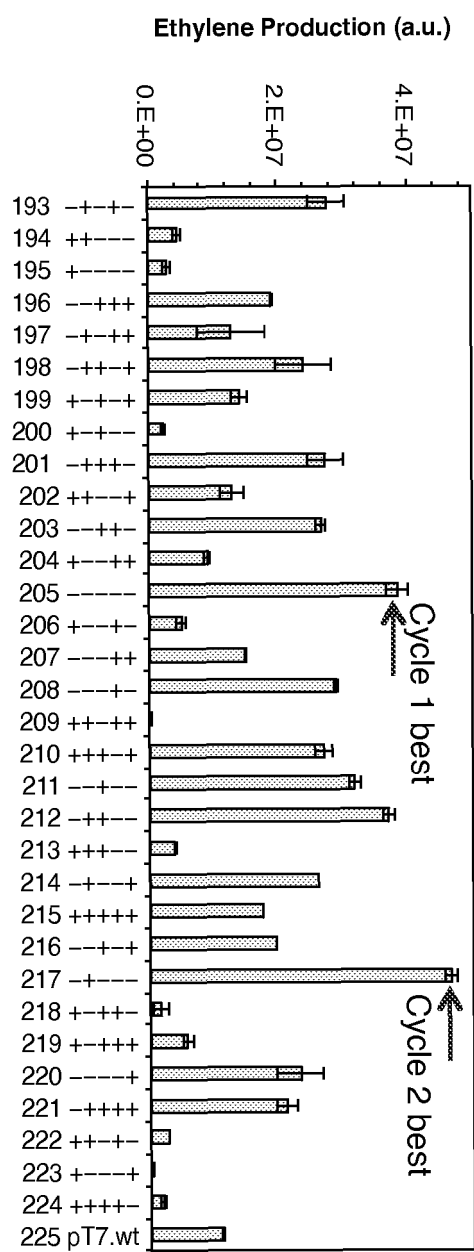
Improved activity (25%) in 2nd DOE engineering cycle

FIGURE 47

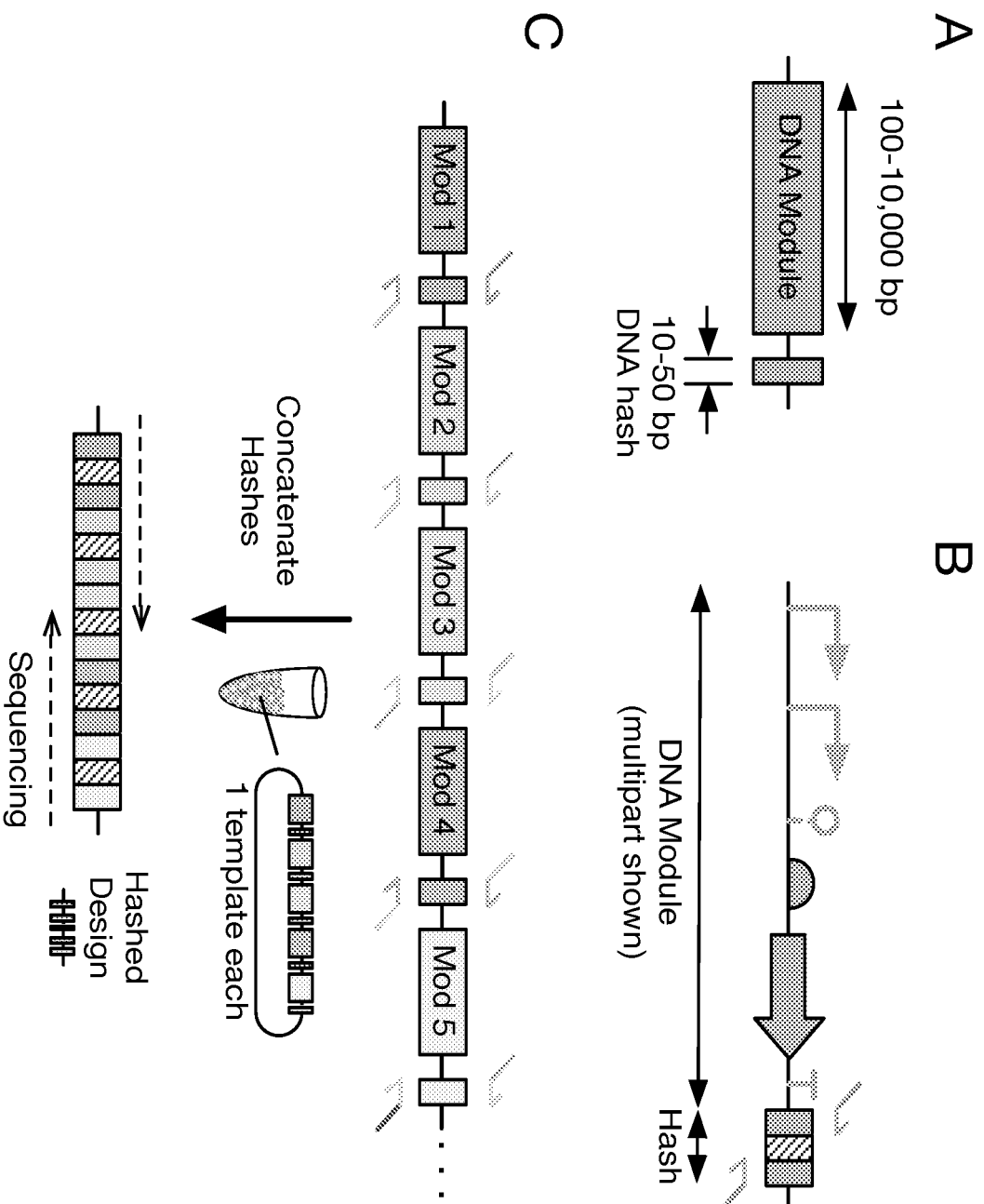


FIGURE 48

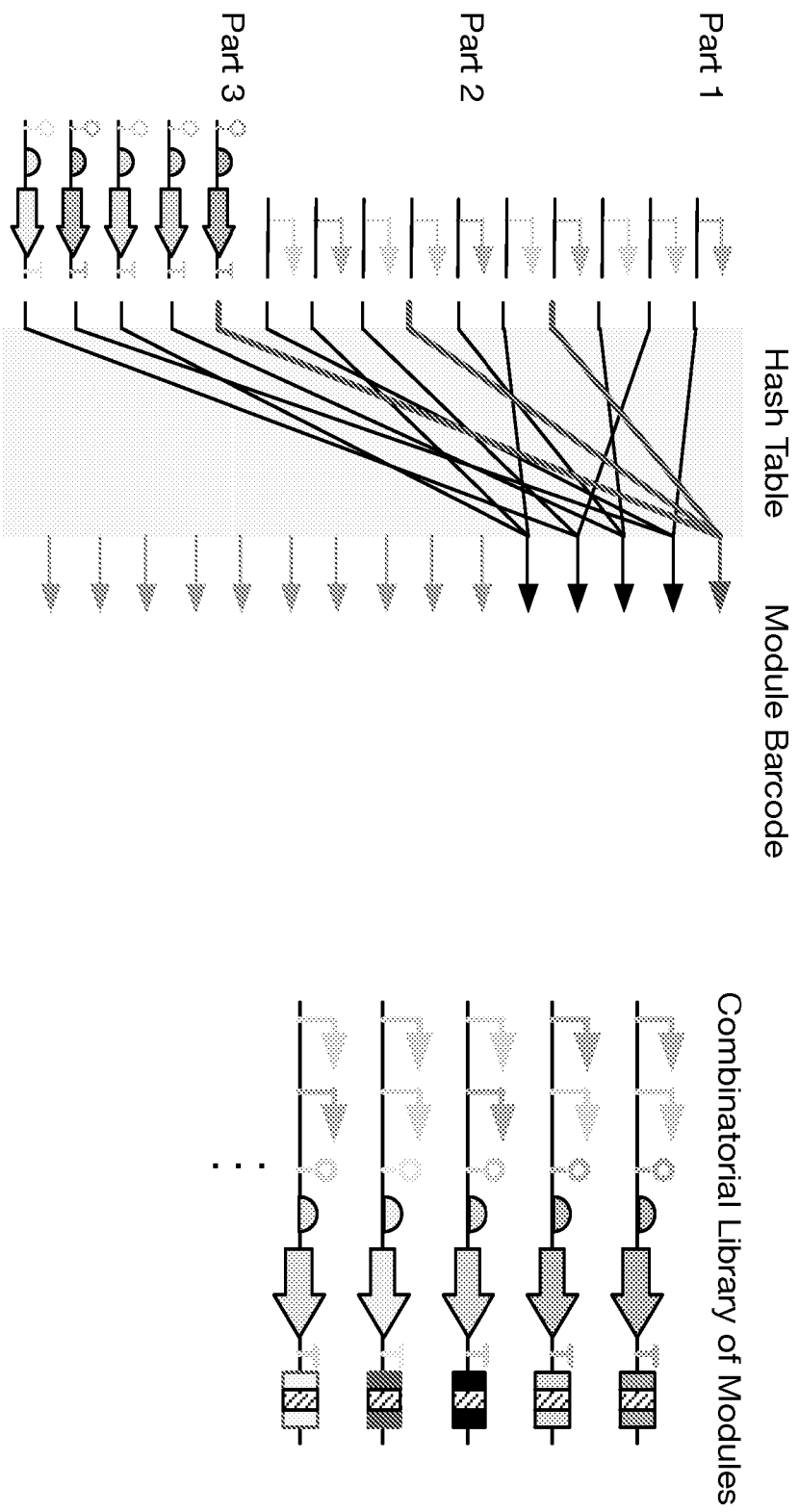


FIGURE 49

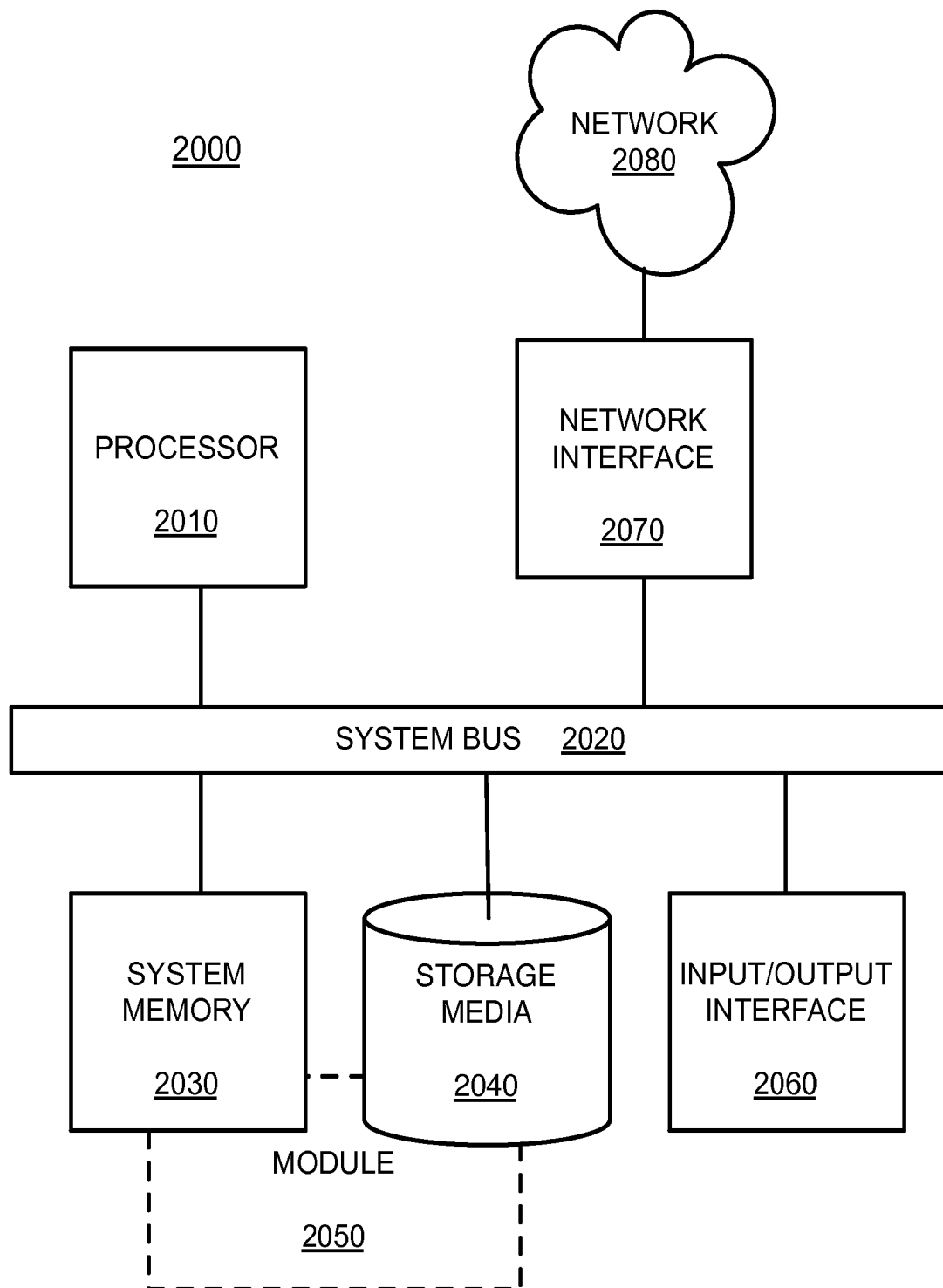


Figure 50

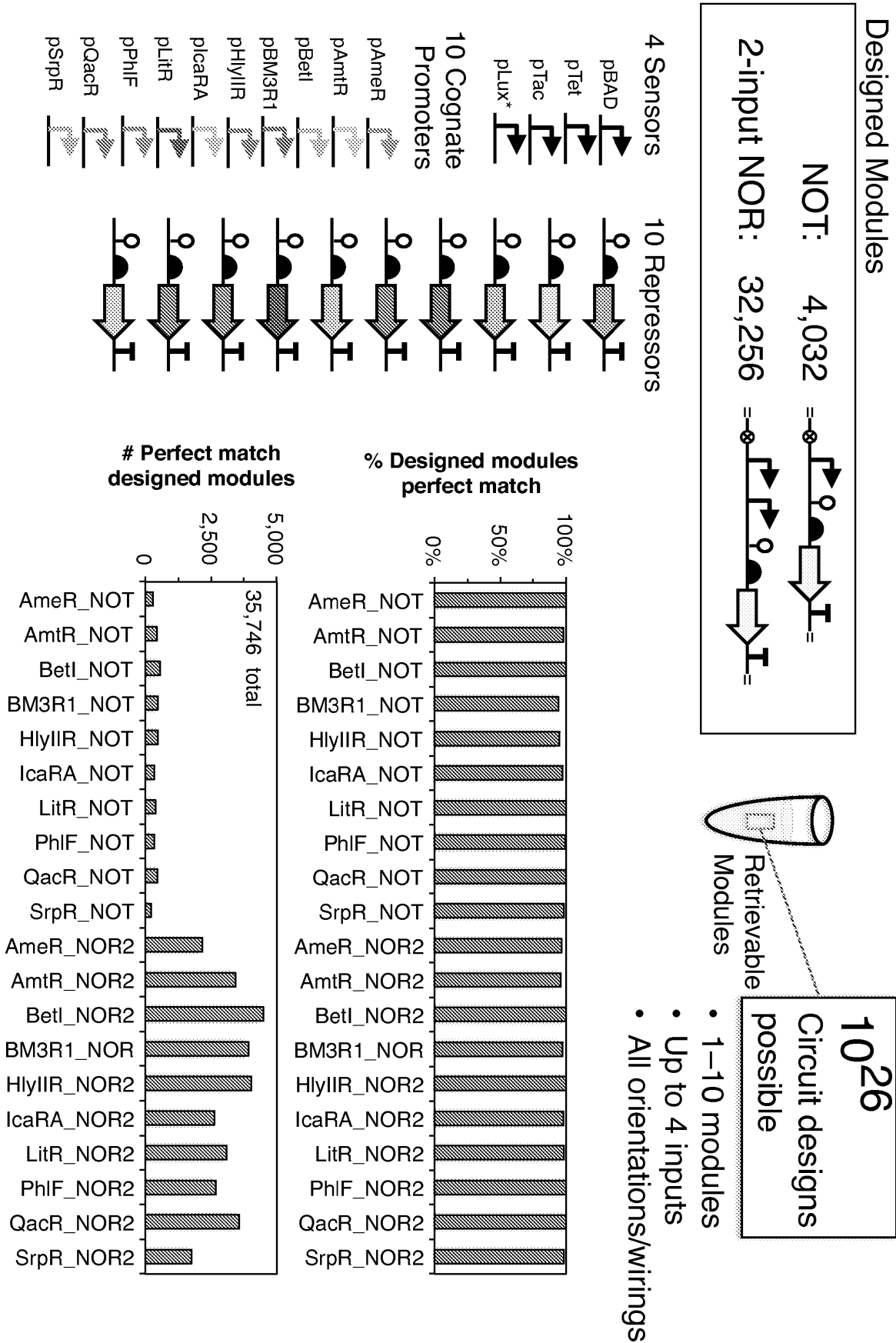
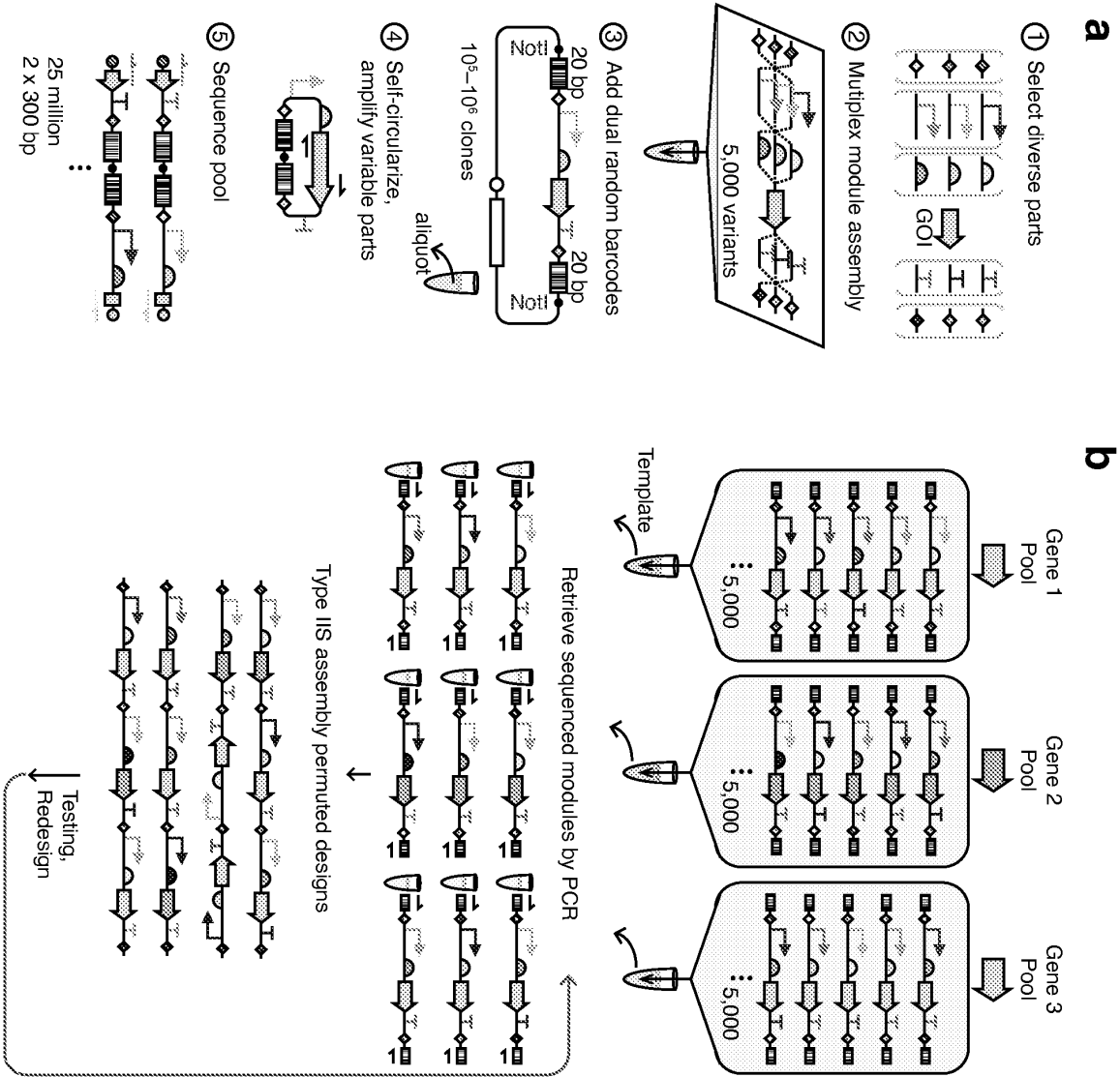




Figure 51

Figure 52



REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Patent documents cited in the description

- US 62003331 [0001]
- US 5744311 A [0023]
- US 6033881 A [0023]
- WO 9001069 A [0023]
- EP 320308 A [0023]
- US 5427930 A [0023]
- US 6027889 A [0023]
- US 6025134 A [0023]
- US 60877812 [0037]
- US 5866336 A, Nazarenko [0042]
- US 20140045728 A1 [0080] [0091]
- US 20120283110 A1 [0080] [0091]

Non-patent literature cited in the description

- **BENJAMIN LEWIN.** Genes IX. Jones and Bartlet, 2008 [0016]
- The Encyclopedia of Molecular Biology. Blackwell Science Ltd, 1994 [0016]
- Molecular Biology and Biotechnology: a Comprehensive Desk Reference. VCH Publishers, Inc, 1995 [0016]
- **VOSKOBOYNIK et al.** *eLife*, 2013, vol. 2, e00569 [0037]
- **SAMBROOK et al.** Molecular Cloning: A Laboratory Manual. Cold Spring Harbor, 1989 [0050]
- **AUSUBEL et al.** Current Protocols in Molecular Biology. Greene Publ. Assoc. & Wiley-Intersciences, 1987 [0050]
- **SMITH ; WATERMAN.** *Adv. Appl. Math.*, 1981, vol. 2, 482 [0054]
- **NEEDLEMAN ; WUNSCH.** *J. Mol. Biol.*, 1970, vol. 48, 443 [0054]
- **PEARSON ; LIPMAN.** *Proc. Natl. Acad. Sci. USA*, 1988, vol. 85, 2444 [0054]
- **HIGGINS ; SHARP.** *Gene*, 1988, vol. 73, 237-44 [0054]
- **HIGGINS ; SHARP.** *CABIOS*, 1989, vol. 5, 151-3 [0054]
- **CORPET et al.** *Nuc. Acids Res.*, 1988, vol. 16, 10881-90 [0054]
- **HUANG et al.** *Computer Appls. in the Biosciences*, 1992, vol. 8, 155-65 [0054]
- **PEARSON et al.** *Meth. Mol. Bio.*, 1994, vol. 24, 307-31 [0054]
- **ALTSCHUL et al.** *J. Mol. Biol.*, 1990, vol. 215, 403-10 [0054]
- **SAMBROOK et al.** Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press, 1989 [0058]
- **SAMBROOK et al.** Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Press, 2001 [0058]
- **AUSUBEL et al.** Current Protocols in Molecular Biology. Greene Publishing Associates, 1992 [0058]
- **AUSUBEL et al.** Short Protocols in Molecular Biology: A Compendium of Methods from Current Protocols in Molecular Biology. Wiley & Sons, 1999 [0058]
- **HARLOW ; LANE.** Antibodies: A Laboratory Manual. Cold Spring Harbor Laboratory Press, 1990 [0058]
- **HARLOW ; LANE.** Using Antibodies: A Laboratory Manual. Cold Spring Harbor Laboratory Press, 1999 [0058]
- **NIELSEN et al.** *Current Opinion in Chemical Biology*, 2013, vol. 17, 878-892 [0066] [0077]
- **ENGLER et al.** *PLoS ONE*, 2008, vol. 3 (11), e3647 [0069] [0126]
- **ELLIS et al.** *Integr. Biol.*, October 2011, vol. 3, 109-118 [0069]
- **SARRION-PERDIGONES et al.** *PLoS ONE*, 2011, vol. 6 (7), e21622 [0069]
- **MERRYMAN ; GIBSON.** *Metabol. Eng.*, 2012, vol. 14, 196-204 [0069]
- **KOSURI et al.** *Nat. Biotechnol.*, 2010, vol. 28 (12), 1295-1299 [0080] [0091]
- **SCHWARTZ et al.** *Nat. Meth.*, 2012, vol. 9, 913-915 [0080] [0091] [0132]
- **CRAIG et al.** *Nat. Methods*, 2008, vol. 5, 887-893 [0087]
- **GERRITS et al.** *Blood*, 2010, vol. 115, 2610-2618 [0087]
- **BERNS et al.** *Nature*, 2004, vol. 428, 431-437 [0087]
- **WEBER et al.** *PLoS ONE*, 2011, vol. 6 (2), e16765 [0102] [0135]
- **KOSURI et al.** *Nat. Biotechnol.*, 2010, vol. 28 (12), 12951299 [0132]